Guangya Wan

Phone: 857-207-5945 | Email: gwan@hsph.harvard.edu
Home Page: https://gwan.netlify.app | Github: https://github.com/wan19990901

Education

New York University (Offer Holder, Expected) *08/2023- 05/2028*
PhD in Data Science (With Research Interest in Deep Learning and Self-Supervised Learning)

Harvard University *08/2021- 05/2023*
Master of Science in Biostatistics (With a Concentration in Data Science, Cross Registered at MIT) **GPA** *3.89/4.00*
Relevant Courses: Advanced Data Science, Machine Learning for Health Care (MIT), Computer Vision (MIT), Parallel Computing, Deep Learning in NLP (MIT), System Programming Development, Distributed Computing, Algorithim

University of Illinois-Urbana, Champaign *09/2017- 12/2020*
Major: *Statistics* (Summa Cum Laude & James Scholar) **Minor**: *Computer Science & Mathematics* **GPA***: 3.94/4.00*
Relevant Courses: Machine learning, Data Mining, Data Structures & Algorithm, Database System, Natural Language Processing, Bioinformatics, Numerical Analysis, Probability & Stochastic Process, Mathematical Statistics, Bayesian Inference

Skills

Programming Languages: Python (Pytorch, Keras, Transformer, lightGBM), SQL(Psql, Snowflake), C++ (OpenMP), R (Rcpp)
Development Toolkits: Git (Github Action), Unix(zsh), VScode, AWS(Sagemaker), GCP(App Engine), Docker, Tex, Markdown

Professional and Research Experience

Chewy **Boston, MA**
Data Science Intern (Pets Health Group)
06/2022-08/2022
- Designed a conjoint analysis to understand consumer's and pets' needs by extracting information from raw text features
- Initiated a data science pipeline to cluster products based on sentence similarity (BERT) of product description and to predict first-90-days demand of pet health products based on the cluster assignment, which outperformed the old methods.

MIT Clinical Machine Learning Group (Machine Learning for Prostate Cancer) **Boston, MA**
Volunteer Research Assistant **Supervisors: by Dr. David Sontag & Madhur Nayan, MD**
01/2022-Now
- Proposed both rule-based methods and BERT-based NLP methods to fill in missing MRI features (Pirads,PSA,etc) from various unstructured reports and wrote well-documented Python scripts and documentations for future use.
- Deployed machine learning model as online calculator to predict lymph node status and performed selection pipeline to select minimum features while persevering the model's macro-F1 score performance.

Boston Children Hospital **Boston, MA**
Research Intern (Computational Health Informatics Program Fellow) ***Supervisor: Dr. William La Cava***
01/2022-Now
- Built fairness evaluation pipeline on both real hospital data and machine learning models on decision related to Emergency Room Admission among different subgroups using demographic parity, differential calibration, and multicalibration
- Applied Medical-BERT from Huggingface to encode the doctors' note as additional test features in fairness measure.
- Proposed a proportional multicalibrated fairness algorithm to postprocess the predicted results of the current machine learning models which control the calibration fairness of a group over intersectional groups without reduced performance.
- Developed reproducible deep learning pipeline to establish the LSTM/CNN baseline performance on the CTU-CHB fecg and toco time series data from MIMIC, applied the same methods on the hospital data (Project 2).

Department of Biostatistics, Harvard School of Public Health **Boston, MA**
Graduate Research Assistant ***Supervisor: Dr. Rui Duan***
09/2021-12/2021
- Implemented Lasso regression algorithm with edited cost function for multiple data integration and federated learning.
- Conducted stimulation study to check the validity of algorithm in Python.

Fields Institute, University of Toronto **Toronto, ON**
The Fields Summer Undergraduate Research Fellow Program (FUSRP) ***Supervisor: Dr. Andreas Hilfinger***
07/2020-09/2020
- Experimented both data-based models (LSTM) and physiological (Biologically inspired) models, read literature to learn about the advantages and drawbacks for both models, and validated result on the real sequential glucose data.
- Selected to receive an award for outstanding presentation at the Canadian Mathematics Conference

National Center for Super Computing Application **Champaign, IL**
Research Assistant Internship (Spin Program) ***Supervisor: Dr Liudmila Mainzer & Dr. Rebecca Lee Smith***
06/2019-12/20
- Implemented data science pipeline (environment config/data cleaning/model fitting/Hyperparameter tuning/postprocess and visulziatize output) to predict the outbreak of West Nile Virus based on weather, and geological covariates, stratified on

social-economical features using parallel computing techniques. Interpret new insights gained from final models (LightGBM) using Gini feature importance and partial dependency plot; achieved a higher macro F1-score than previous papers.
- Conducted survival analysis on campus-survey Cov-19 data, achieved a great C-index score by combining machine learning with the classical methods, and wrote a report to the campus policer makers. (Project 2)

Relevant Project

Subjectivity Analysis of Clinical Notes (MIT EECS 6.S982 Clinical Data Learning and Deployment Research Project)
08/2022-Now
- Built cohorts of interest (age/gender/smoking, etc) along with the doctor's notes from admission from MIMIC dataset.
- Investigated pre-trained subjectivity and bias detection BERT model and fine-tuned the model on a subset of clinical text with manually label provided by collaborators and make inference with the model on all cohort text data to get the bias label.
- Performed statistical analysis on inferred data to detect systematic biases against certain subgroups from doctors' notes.

Automatic Differentiation and Optimization Package Development (Harvard CS 207 System Development Final Project)
08/2022-Now
- Developed a Python package that implements dual number and forward mode for automatic differentiation which was used to calculate the Jacobian required for optimization and backpropagation submodule (SGD,NADAM/RMSprop/BFGS,etc).
- Validated code by writing an extensive test suite using Python's unit test pipeline (pytest, docstring test, Github Action, 100% code coverage reflected**)** with detailed documentation, published code on git, and distributed the package via PyPI.

Enhance Pretrained Model with Multiple Knowledge Integration (MIT EECS 6.861 Deep Learning in NLP Research Project)
08/2022-Now
- Established baseline BERT model for multi-labelled ICD code prediction on multiple disease diagnosis.
- Proposed a special BERT representation which combined knowledge from different source (medical paper, wiki pedia of various disease) to incorporate the knowledge about disease, risk factors, and symptom to increase the generalizability of model and outperformed the previous model.

Parallel Collaborative Filtering Recommendation (Harvard CS 205 High Performance Computing Final Project)
01/2022-05/2022
- Parallel web scraped movie review data with python's multi-processing and beautifulsoup saved data using with MongoDB.
- Programmed similarity measure function, sort function, and read/write function with C++ (OpenMP) to parallelize the code
- Achieved a 10x boost in terms of benchmark speed compared to unparallel C++ implementation.

Honors and Rewards

Dean's List	Fall 2017 - Spring 2020
Edmund J. James Scholar	Spring 2018 - Spring 2020
Joint Mathematical Meeting Poster Presentation, Honorable Mention	Jan 2020
Student Pushing Innovation (SPIN) Fellowship	June 2020
Canadian Undergraduate Mathematical Conference, Best Presentation Nominee	Aug 2020
Undergraduate Research Certificate	May 2020
Phi Beta Kappa Honor Society Nomination	Dec 2020
Computational Health Informatic Program (CHIP) Fellowship	Jan 2022

Conferences and Publications

Guangya Wan, Hao Wang, Ronglin Wang, "Statistical Analysis of Weather Forecasts", Joint Mathematics Meetings. **(Honorable Mentions)** ***Denver, CO, Jan 17th , 2020***
Guangya Wan, Shubham Rawlani, "Risk of West Nile Virus among Chicago area",CSL Conference ***Urbana, IL, Feb. 28th, 2020***
Guangya Wan, Lily Yang, Preetham Bachina, Ruo Ning Qiu, "Testing Mathematical Models of Diabetes Against Blood Glucose Data" Canadian Undergraduate Mathematical Conference (**Best Presentation Nominee**), ***Toronto, Aug. 24th, 2020***
Guangya Wan, Xiaorui Wang, "The Iterated Prisoner's Dilemma " , Joint Mathematics Meetings, ***Remote, Jan 17th , 2021***
Guangya Wan, William La Cava, Elle Lett, "Proportioanl Multicalibration" (CHIL Preprint, https://arxiv.org/abs/2209.14613)
Guangya Wan,Weihao Ge, John Uelmen, Liudmila Sergeevna Mainzer, Rebecca Lee Smith, "Stratified random forest model to identify human West Nile Virus infection risk factor in Chicago" (PLOS One, under revision)

Teaching Experiences

Course Assistant (Harvard CS109 B, Advaned Data Science II) 01/2023-Now
- Helped design and grade the homework; hosted office hour one hour a week to answer students' questions

Research Mentor (National Center for Super Computing Application, Genomics Group) *09/2020-12/2020*
- Lead five local high school Students to study and run basic machine learning code using python on real genomics dataset.

Statistics Tutor (Department of Statistics, University of Illinois, Urbana Champaign) *09/2019-12/2019*
- Arranged 1 on 1 meeting with statistics major student on general statistics understanding and registration advising.

Activities and Leadership

- Lead Members of Harvard Astrophotography Club (STAHR) (Help organize workshop to teach students how to take a photo of galaxy/nebula; participate in monthly dark-sky trip)
- Members of Harvard Poker Club (participate weekly poker games hosted by Citedal)
- Members of Harvard Hiking Club(participate in monthly back-country hiking trip in New England)
- Lead Members of HackIllinois Hackathon (Ameliorated homework system "Prairielearn" for the University of Illinois)
- Lead Members of Illinois Statistics Club (Help organize statistics career development events and R programming camp)