Yuan Lu

Homepage: https://yuan-lu-bryan.github.io/
Email: yul331@ucsd.edu
Github: yuan-lu-bryan

Education

University of California San Diego — La Jolla, CA, USA
PhD in Data Science (Machine Learning) — *Sep 2024 - June 2029*

- ***Supervisor:*** *Professor Yusu Wang*
- ***Research Interests:*** *Geometric and Topological Deep Learning, Graph Neural Networks, Gaussian Processes and Bayesian Optimization on Graphs, Deep Generative Models, Machine Learning for Computational Biology*

University of Cambridge — Cambridge, UK
MPhil in Machine Learning and Machine Intelligence — *Oct 2022 - Sept 2023*

- ***Grade****: Pass, 73.36/100*
- ***Focused Areas:*** *Bayesian Machine Learning, Computer Vision, Deep Learning and Structured Data, Graph Neural Networks, Speech and Language Processing, Reinforcement Learning*

University College London — London, UK
BSc (Hons) in Statistics — *Sept 2019 - June 2022*

- ***Grade****: First Class Honors, 82.14/100 (Top 5% in cohort)*
- ***Selected Coursework:*** *Complex Analysis (92/100), Machine Learning for Domain Specialists (95/100), Statistical Inference (86/100), Stochastic Processes (86/100), Stochastic Methods in Finance (88/100)*

Publications

Yuan Lu, Haitz Sáez de Ocáriz Borde, Pietro Liò. AMES: A Differentiable Embedding Space Selection Framework for Latent Graph Inference. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023.

Research Experience

Research Assistant in Geometric Deep Learning — **University of Cambridge**
Advisor: Professor Pietro Liò, Haitz Sáez de Ocáriz Borde — *Mar 2023 - Nov 2023*

- Worked on Latent Graph Inference with Professor Pietro Liò from the Department of Computer Science and Technology, University of Cambridge and Haitz Sáez de Ocáriz Borde from the University of Oxford
- Improved the robustness of Manifold-based Latent Graph Inference modules via self-attention-based weighted combinations of latent graphs from different embedding spaces
- Developed a novel framework for automatically selecting appropriate embedding spaces for different downstream GNN tasks via an attention-based gradient descent, whose effectiveness has been validated by a gradient-based interpretability analysis
- Published a paper in NeurIPS 2023 **Symmetry and Geometry in Neural Representations Workshop**

Research Assistant in Bayesian Optimization — **University College London**
Advisor: Dr François-Xavier Briol — *Sept 2021 - April 2022*

- Compared and contrasted GP-UCB (a non-Robust Bayesian Optimization algorithm) and StableOpt (a Robust Bayesian Optimization algorithm) on synthetic toy problem to validate the effectiveness of StableOpt on robust optimization problems
- Reframed an offshore wind farm layout optimization problem as a perturbed optimization problem and successfully applied StableOpt to it, achieving promising results
- Coded extensively in Python with NumPy, SciPy and Skopt for Bayesian Optimization algorithms implementation and all synthetic and real-world experiments
- Obtained high distinction for the report titled **Robust Wind Farm Layout via Bayesian Optimization**

Selected Projects

Conditional Neural Processes: A Comparison Study
Team Paper Replication Project — *Jan 2023 - Mar 2023*

- Implemented Vanilla Conditional Neural Processes, Attentive Conditional Neural Processes and Convolutional Conditional Neural Processes from scratch with PyTorch
- Reproduced the 1D function regression tasks on synthetic data and 2D image completion tasks on MNIST, CelebA and Multi-digit MNIST, achieving better results than those reported in the original Conditional Neural Processes paper
- GitHub repository: *https://github.com/tlc4418/neural-processes*

Reinforcement Learning Algorithms for Grid-World Environments
Individual Course Project *Feb 2023 - Mar 2023*

- Implemented Value Iteration, Policy Iteration, SARSA, Expected SARSA and Q-learning Reinforcement Learning algorithms from scratch using standard Python packages in discrete grid-world environments
- Plotted cumulative rewards and learned policies for SARSA and Q-learning with different hyperparameters to compare their learning behaviour and performance differences

TIMIT/LibriSpeech Speech Recognition Using Foundation Models
Individual Course Project *Feb 2023 - Mar 2023*

- Fine-tuned a Wav2vec2-based foundational model on TIMIT and LibriSpeech dataset with PyTorch and Hugging Face, achieving a state-of-the-art Word Error Rate (WER) of 10% on LibriSpeech test set with only ten epochs
- Fine-tuned another WavLM-based foundation model on LibriSpeech, improving the WER further to 8% on LibriSpeech test set without language model fusion

A Keyword Spotting System for Swahili Language
Individual Course Project *Feb 2022 - Mar 2023*

- Built a keyword spotting (KWS) system for a low-resource language, Swahili, based on the 1-best output of an automatic speech recognition (ASR) model with posterior scores, using morphological decomposition and grapheme confusion matrix techniques to address the issue of out-of-vocabulary (OOV) queries
- Implemented score normalisation and system combination methods to improve the overall performance of the KWS system measured by maximum term weighted value (MTWV)

Latent Dirichlet Allocation for Topic Modelling
Individual Course Project *Nov 2022 - Dec 2023*

- Trained a Bayesian Mixture of Multinomials model and a Latent Dirichlet Allocation model via Gibbs sampling on 2000 documents
- Compared the performances of the two models via per-word perplexity and word entropy.

Probabilistic Ranking for 2011 ATP Men's Tennis Singles Data
Individual Course Project *Nov 2022 - Nov 2023*

- Estimated the probabilistic rankings of the skills of 107 tennis players using the 2011 ATP men's tennis singles data under the TrueSkill Ranking System
- Used Gibbs sampling and Expectation Propagation to approximate the posterior skill distributions of the 107 players

Work Experience

Data Science Summer Intern **Johnson & Johnson China, Shanghai**
Biometrics Q & S Team *July 2021 - Sept 2021*

- Developed an automation tool in R that can automatically extract tables from Clinical Summary Report Synopsis pdf, label each paragraph with correct content tags, and output a fully structured data frame of texts
- Developed an R Shiny app that implements my automation tool in a user-friendly manner and generates a maintainable database of labelled texts and extracted tables from Clinical Summary Report pdf
- Developed another R Shiny app that supports advanced statistical analysis, such as data filtration, cross-table generation, summary statistics calculation and data visualization of biostatistics data for the team.

Honors & Awards

- **First Year PhD Fellowship**, Halıcıoğlu Data Science Institute, UC San Diego, 2024-2025
- **Dean's List**, UCL Faculty of Mathematical & Physical Sciences, 2022
- **First Year Sessional Prize**, UCL Department of Statistical Science, 2020

Skills

Programming Languages: Python, MATLAB, R, SAS, LaTeX
Frameworks & Tools: PyTorch, PyTorch Geometric, TensorFlow, Scikit-learn, Hugging Face, Unix
(Human) Languages: Mandarin Chinese (native), English (full professional proficiency)

Other Activities

Discover Citadel LLC/Citadel Securities Spring Insight **Citadel LLC, London**
Quantitative Research Team *Apr 2020 - Apr 2020*

- Gained deep insights into the main functions of the Quantitative Research and Trading department.
- Attended a one-to-one mock interview with one of the Quantitative researchers and developed a further understanding of the Quantitative Research industry.