CHUFAN GAO

219-239-8008 ⋄ gaoandy1445@gmail.com ⋄ Urbana, United States
LinkedIn: chufangao ⋄ Github: chufangao ⋄ Google Scholar: rBlZICgAAAAJ

Research Statement: I am currently a PhD Student at University of Illinois Urbana-Champaign advised by Professor Jimeng Sun. My areas of focus include Natural Language Processing, Sequential Event Data, and Machine Learning for Healthcare in general. I am also broadly interested in time-series data and less than supervised Machine Learning.

EDUCATION

PhD in Computer Science August 2022 – 2026
University of Illinois Urbana-Champaign *Urbana, Illinois*

- PhD Student advised by Professor Jimeng Sun (GPA 4.0). Working in Natural Language Processing, Machine Learning for Healthcare, Clinical Trial Outcome Extraction
- Relevant Courses – Advanced NLP, Text Mining, Deep Learning for Healthcare, Advanced Information Retrieval

Masters of Science in Robotics August 2020 – August 2022
Carnegie Mellon University *Pittsburgh, PA*

- School of Computer Science, Research Masters (GPA: 3.90), Thesis: Addressing Time-series Signal Quality in Healthcare Data
- Relevant Courses - Math Fundamentals for Robotics, Computer Vision, Probabalistic Graphical Models, Machine Learning, Convex optimization

Bachelor of Science in Computer Science and Mathematical Statistics August 2016 – May 2019
Purdue University *West Lafeyette, IN*

- **With Honors and Distinction** (GPA: 3.90)
- Relevant Courses (** indicates graduate level*) – Machine Learning*, Algorithms*, AI*, Graphical Models, Data Structures and Algorithms, Advanced Linear Algebra, Differential Equations, Real Analysis, Probability*, Statistical Theory*

WORK EXPERIENCE

Research May 2023 – August 2023
Medidata Solutions (Dassault Systmes) *New York City, NY*

- Summer Research Project: `HawkesVAE`: Sequential Patient Event Synthesis For Clinical Trials (In Submission)

Research Intern June 2022 – August 2022
IQVIA Analytics Center of Excellence *Remote*

- Conducted 2 thorough reviews regarding Insilico Clinicial Trials and Machine Learning for Clinical Trials

Research Associate August 2019 – August 2022
Carnegie Mellon University Robotics Institute *Pittsburgh, PA*

- Conducted various research projects in partnership with the AutonLab and University of Pittsburgh Medical Center (Advised by Professor Artur Dubrawski)
- Published 1 accepted paper in Neurips (ML4H) Workshop
- Published 2 accepted student abstracts in AAAI Student Track, 1 accepted paper in AAAI Symposium–Artificial Intelligence for Predictive Maintenance
- Published 2 medical abstracts in American American Thoracic Society

Robotics Institute Summer Scholar (RISS) June 2019 – August 2019
Carnegie Mellon University Robotics Institute *Pittsburgh, PA*

- Robotics Institute Summer Scholar (RISS) program (2-3% acceptance rate) - Investigated methods into detecting physiological state changes via deep unsupervised learning mentored by Professor Artur Dubrawski
- Methods include a custom Pytorch implementation of dilated CNNs for sequence embedding and autoencoders with attention
- Resulted in acceptance to NeurIPS ML4H workshop as well as a staff research position

NSF REU Undergraduate Researcher June 2018 – August 2018
DePaul University College of Computing and Digital Media *Chicago, IL*

- Medix REU Program (<10% acceptance rate) - Implemented a custom 3D Generative Adversarial Networks and 3D CNN to improve performance of Computer-Aided Detection systems under Professor Jacob Furst and Professor Daniela Raicu
- Resulted in oral presentation and publication of *Augmenting LIDC dataset using 3D generative adversarial networks to improve lung nodule detection* in SPIE Medical Imaging conference

Undergraduate Researcher May 2017 – May 2018
Purdue University *West Lafayette, IN*

- Student Learning Research - Conducted statistical analysis on the effects of active learning classes on future student performance Mentored by Professor Clarence Maybee. Presented *Impact of Active Learning on Future Student Performance* via invited oral presentation in Purdue Journal of Undergraduate Research.
- Deep Health Metric Prediction - Acquired experimental data, automated preprocessing workflow, and wrote a program to determine health metrics such as Heart Rate, HRV, and PPG from videos using DeepFace mentored by Professor Vaneet Aggarwal

ACADEMIC PROJECTS

PCALG in Python: Read respective papers on 4 Causal learning algorithms (PC, FCI, GES, LINGAM) and ported over R and C++ code from pcalg (R Causal Inference Package) to Python.

Weak Text Classification: Combined Data Programming (Snorkel) and glove word similarities to extract useful keywords for imdb reviews text classification. Resulted in a technical report: ”The word is mightier than the label: Learning without pointillistic labels using data programming”.

Time Series Anomaly Detection: Extended an existing AAAI paper's graph neural network to account for uncertainty in estimates. Resulted in a technical report: ”Learning graph neural networks for multivariate time series anomaly detection”.

TEACHING AND MENTORSHIP

Teaching Assistant January 2024 – May 2024
University of Illinois Urbana-Champaign *Urbana, IL*

- Created, graded, and reviewed labs, projects, and tests for CS598 Deep Learning for Healthcare.

Veritas AI Mentor Spring 2022
Veritas AI *Remote*

- Lead and mentored multiple groups of high school students over a 10-week time period to learn machine learning and classify CIFAR-10 images using a CNN. Github Link.

AI4ALL Mentor Summer 2021
Carnegie Mellon University *Pittsburgh, PA*

- Advised 5 high school students one-on-one over a 2-week time period to use transfer learning and DenseNet to achieve over 90% accuacy in plant disease classification.
- Created and tested project template code in Google Collaboratory, ensuring that memory usage was as low as possible for the above task.

HONORS AND AWARDS

Scholarships and Academic Awards

Boeing Scholarship - disbursed based on academic merit in CS. 5/1900 CS students at Purdue.

- Purdue Presidential Scholarship - disbursed based on high academic achievement; leadership and service in school/community. 830/40,000 students at Purdue.
- Gordan L. Walker Scholarship - disbursed based on continuing academic achievement in mathematics. 1 out of all Math students at Purdue.
- Purdue West Lafayette Deans List (all years), Honors College Member (all years).

DJI Drone Challenge Summer 2019

- Led a team of 5 in a drone challenge following a path specified by aruco tags autonomously. Implemented functionality of viewing AR holographic images through the drone camera. Created android app to switch between drone modes.
- 1st place out of 8 teams and 40 competitors

UBTech Humanoid Challenge Summer 2019

- Led a team of 6 in programming and teleoperating a humanoid robot in ROSpy with a raspberry pi that could effectively grasp and move a small object.
- 1st place out of 6 teams and 40 competitors.

PUBLICATIONS (* DENOTES EQUAL CONTRIBUTION)

1. Hanyin, W., **C. Gao**, C. Dantona, B. Hull, J. Sun ”DRG-LLaMA: tuning LLaMA model to predict diagnosis related group for hospitalized patient,” in Nature Digitial Medicine, 2024
2. **C. Gao**, N. Gisolfi, and A. Dubrawski, ”Signal quality auditing for time-series data,” in AAAI Fall Symposium: Artificial Intelligence for Predictive Maintenance, 2022, Code
3. **C. Gao**, ”Addressing time-series signal quality in healthcare data,” Masters thesis, Carnegie Mellon University, Link
4. **C. Gao***, M. Goswami*, J. Chen, and A. Dubrawski, ”Classifying unstructured clinical notes via automatic weak supervision,” in Machine Learning for Healthcare, 2022. Code Link
5. J. H. Yoon, **C. Gao**, J. Kim, J. H. Kim, T. Lagattuta, S. Helman, M. Hravnak, M. R. Pinsky, and G. Clermont, ”Prediction of hypovolemic instability in normal volunteer blood donors using machine learning (Abstract)”, in American Thoracic Society, 2022 Link
6. **C. Gao**, A. Dubrawski, M. Pinsky, G. Clermont, and J. Yoon, ”Identification and explanation of severity of bleeding-induced hypovolemia using unsupervised deep learning (Abstract)”, in American Thoracic Society, 2021 Link
7. M. Goswami, L. Chen, **C. Gao**, and A. Dubrawski, ”Modeling involuntary dynamic behaviors to support intelligent tutoring (Student Abstract),” in AAAI Link
8. S. Peng, L. Chen, **C. Gao**, and R. J. Tong, ”Predicting students attention level with interpretable facial and head dynamic features in an online tutoring system (Student Abstract)” in AAAI Link
9. **C. Gao**, F. Falck, M. Goswami, A. Wertz, M. R. Pinsky, and A. Dubrawski, ”Detecting patterns of physiological response to hemodynamic stress via unsupervised deep learning,” in NeurIPS ML4H Workshop, 2019 Link
10. **C. Gao**, S. Clark, J. Furst, and D. Raicu, ”Augmenting LIDC dataset using 3d generative adversarial networks to improve lung nodule detection,” in SPIE Medical Imaging, 2019 Link
11. **C. Gao**, ”Out of the box: Impact of active learning on future student performance,” The Journal of Purdue Undergraduate Research, 2018 Link

Technical Reports

1. **Gao, C.**, Fan, X., Sun, J., Wang, X. (2023). ”PromptRE: Weakly-Supervised Document-Level Relation Extraction via Prompting-Based Data Programming,” 2023 Link
2. Z. Wang* **C. Gao*** J. Sun ”MediTAB: Scaling Medical Tabular Data Predictors via Data Consolidations, Enrichment, and Refinement,” 2023 Link
3. Z. Wang* **C. Gao*** J. Sun ”A Survey: In Silico Trials,” 2023 Link
4. M. Goswami*, **C. Gao***, B. Boecking, A. Dubrawski, ”Active Learning for Weakly Supervised Model Refinement,” 2022 Link
5. S. Ray*, S. Lakdawala*, M. Goswami*, and **C. Gao***, ”Learning graph neural networks for multivariate time series anomaly detection”, 2022 Link
6. **C. Gao*** and M. Goswami*, ”The word is mightier than the label: Learning without pointillistic labels using data programming”, 2021 Link

In Progress

1. Z. Wang*, **C. Gao***, J. Sun ”Meditab: Healthcare Tabular Prediction with LLMs” (In Submission)
2. **C. Gao**, M. Beigi, A. Shafquat, J. Sun ”Hawkes-Process Variational Autoencoder for Generating Synthetic Clinical Trials” (In Submission)

TECHNICAL STRENGTHS

Programming Languages – Python

Frameworks / Tools – Pytorch, Sklearn, Tensorflow Keras

ADDITIONAL PROJECTS AND SERVICE

Committees

- Carnegie Mellon University Robotics Institute Summer Scholars (RISS) Admissions Committee (2020-2022): Reviewed applicants on quality of fit to RISS. Produced forms and documentation used to streamline application process.

Reviewer Duties

- International Conference on Learning Representations (ICLR), 2022, 2024
- NeurIPS, 2019-2021
- ACM Conference on Health, Inference, and Learning (CHIL), 2020