CHUFAN GAO

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

Research Statement: I am currently a PhD Candidate at the University of Illinois Urbana-Champaign advised by Professor Jimeng Sun. My areas of focus include information extraction, tabular data, and NLP for healthcare. I am also broadly interested in less-than-supervised ML and memory-augmented ML.

EDUCATION

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

- PhD Candidate 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 Advised by Professor Artur Dubrawski (GPA: 3.9), Thesis: Addressing Time-series Signal Quality in Healthcare Data
- Relevant Courses - Math Fundamentals for Robotics, Computer Vision, Probabilistic Graphical Models, Machine Learning, Convex optimization

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

- **With Honors and Distinction** (GPA: 3.9)
- 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

NLP Research Intern June 2024 – August 2024
Optum (United Health Group) *Remote*

- Synthetic biomedical NER dataset generation, submitted to NAACL. Additionally, submitted 2 patents relating to biomedical NER dataset generation as well as memory-augmented biomedical NER.

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

- Sequential clinical trial patient event synthesis. Collaborated on existing research that resulted in an ICML best workshop paper

Research Staff 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, 2 accepted student abstracts in AAAI Student Track, 1 accepted paper in AAAI Symposium–Artificial Intelligence for Predictive Maintenance, 2 medical abstracts in American American Thoracic Society

Research Intern June 2019 – August 2019
Carnegie Mellon University Robotics Institute *Pittsburgh, PA*

- Part of the 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. Created a custom Pytorch implementation of dilated CNNs for sequence embedding and autoencoders with attention and Resulted in acceptance to NeurIPS ML4H workshop as well as a staff research position

NSF 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

SELECTED PUBLICATIONS (* DENOTES EQUAL CONTRIBUTION)

1. **C. Gao**, X. Wang*, J. Sun*, "TTM-RE: Memory-Augmented Document-Level Relation Extraction" in ACL (Main) 2024 Code Link
2. Z. Wang*, **C. Gao***, J. Sun "Meditab: Healthcare Tabular Prediction with LLMs," in IJCAI 2024 Code Link
3. Hanyin, W., **C. Gao**, C. Dantona, B. Hull, J. Sun "DRG-LLaMA: tuning LLaMA model to predict diagnosis related group for hospitalized patients," in Nature Digitial Medicine, 2024 Code Link
4. **C. Gao**, "Addressing time-series signal quality in healthcare data," Masters thesis, Carnegie Mellon University, 2022 Code Link
5. **C. Gao***, M. Goswami*, J. Chen, and A. Dubrawski, "Classifying unstructured clinical notes via automatic weak supervision," in Machine Learning for Healthcare (MLHC), 2022. Code Link
6. **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 Code Link

Technical Reports

1. **Gao, C.***, Pradeepkumar, J., Das, T., Thati S., Sun, J. (2024). "Automatically Labeling $200B Life-Saving Datasets: A Large Clinical Trial Outcome Benchmark," 2024 Code Link
2. **Gao, C.**, Fan, X., Sun, J., Wang, X. (2023). "PromptRE: Weakly-Supervised Document-Level Relation Extraction via Prompting-Based Data Programming," 2023 Link
3. Z. Wang* **C. Gao*** J. Sun "A Survey: In Silico Trials," 2023 Link

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 an 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.

TEACHING AND MENTORSHIP

Teaching Assistant for CS598 Deep Learning for Healthcare January 2024 – May 2024
University of Illinois Urbana-Champaign *Urbana, IL*

- Created, graded, and reviewed labs, projects, and tests. Top answerer on Piazza.

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% accuracy in plant disease classification. Created and optimized project template code to fit Google Collab memory requirements.

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 the application process.
- Sunstella Summer Camp 2023: Mentored 3 participants in ML projects, one of which went on to become a PhD Student at UIUC

Reviewer Duties

- ICLR 2022, 2024, 2025
- NAACL 2024
- ACM KDD 2024
- NeurIPS 2019-2021, 2024
- ACM CHIL 2020