Benjamin Irving

irving.b@northeastern.edu | Website: biirving.github.io | 510-290-1513 | Google Scholar | GitHub: biirving
Available: May 2025

Education

Northeastern University, Khoury College of Computer Sciences 2020-2024
Bachelor of Science in Computer Sciences, minor in Physics GPA: 3.7/4.0 | Dean's List

Relevant Courses: Group Theory | Reinforcement Learning | Machine Learning and Data Mining 2 | Robotic Science and Systems | Computer Systems | Physics for Engineering 1 | Software Engineering

Industry Experience

Software Engineering Intern, Vecml September 2024 – Present

- Building Retrieval Augmented Generation (RAG) systems on the edge
- Implementing Table RAG and Knowledge Graphs

Machine Learning SDE Intern, Amazon Web Services, Annapurna Labs June 2022 – Aug 2022
Cupertino, California

- Implemented Transformer models (BEiT, ViT) using PyTorch and Neuron SDK for EC2 Inf.2x instances
- Achieved 83.96% accuracy on ImageNet validation with BEiT and 76.92% with ViT
- Created extendable repository for AWS-Neuron team, benchmarking throughput, latency, and cost

Quality Engineer Intern, Optum (UnitedHealth Group) Jan 2022 - June 2022
Boston, Massachusetts

- Automated test suite using Java, Maven, and Selenium, reducing test workload from 96 to 4 hours
- Implemented test automation for Optum Financial (1.8M customers, $200M annual claims)

Publications

1. Irving, B., & Schoene, A.M. (2024). "MEANT: Multimodal Encoder for Antecedent Information." In *Proceedings of EMNLP 2024*, pages 8579-8600.
2. Zevallos, R. J., Ortega, J. E., & Irving, B. (2024). "Related Work Is All You Need." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 13874-13878.
3. Schoene, A.M., Garverich, S., Ibrahim, I., et al. (2024). "Automatically extracting social determinants of health for suicide: a narrative literature review." *npj Mental Health Research*, 3(51). doi:10.1038/s44184-024-00087-6

Research Experience

Research Assistant, Institute for Experiential AI Oct 2023 – Present

- Training large language models on mental-health data using multiple GPU nodes
- Implementing extensible NER library supporting nested-NER capabilities
- Building infrastructure with PyTorch, Cython, Triton, OpenMPI, and HuggingFace libraries

Research Assistant, Helping Hands Lab April 2024 – June 2024

- Worked with Offline Meta Learning and diffusion.
- Trained deep reinforcement learning models to complete tasks with a robotic arm, working with equivariance and Proximal-Policy Optimization (PPO) to improve sample efficiency.
- Used ross and C++ to run on real robots

Undergraduate Researcher, JSALT NLP Workshop, Johns Hopkins University June 2023 – Aug 2023
Le Mans, France

- Developed large-scale graph embedding procedures for 200M+ nodes using randomized SVD and Chebyshev iterations
- Reduced runtime by 24% and memory requirements by 25% (200GB) through optimized implementation
- Implemented solutions in C, Python, and Cython for maximum performance
- Project: `github.com/kwchurch/JSALT_Better_Together`