ILEANA RUGINA

(617)655 4815 | irugina.github.io | ileana.rugina.2@gmail.com

WORK AND RESEARCH EXPERIENCE

Software Engineer at Emerald Innovations
Cambridge, MA — Oct. 2023 - Present
Developed ML systems at digital health MIT CSAIL spinoff to extract health biomarkers from RF signals:
- Refined interactive dashboards to visualize ML predictions and RF signal patterns, enabling faster model iteration.
- Deployed a manifest system to track data files uploaded asynchronously by IoT sensors, ensuring data integrity.
- Reduced ML inference costs and latencies by migrating to spot cloud instances, optimizing the job scheduling algorithm, as well as optimizing SQL queries through index strategies and schema redesign.
- Implemented secure model serving infrastructure using Terraform, automating SSL/TLS certificate management.

Machine Learning Software Engineer at Gridspace
Los Angeles, CA — Sep. 2021 - Oct. 2023
Developed speech and language AI systems at conversational AI startup (Stanford/SRI Speech Labs collaboration)
- *Speech AI:* Developed and deployed a speech synthesis model with controllable prosody using variational embeddings to improve virtual agent capabilities. Improved performance of a dual-mode conformer model with joint CTC/RNN-T decoding to eliminate 85% of hallucinations.
- *LLM Inference:* Implemented speculative sampling to lower encoder-decoder summarization latency by up to 40%.
- *Dialog Systems:* Improved planning algorithm and integrated vector embeddings in dialog system.
- *Platform:* Coordinated data collection efforts, interacting with both in-house and third-party teams for audio transcription and LLM instruction tuning. Improved reliability and accuracy of high-level analytics through distributed system optimization and anomaly detection.
- *Other Responsibilities:* production deployments, mentored interns, polished demos, recruiting.

MEng Research Assistant at MIT Soljačić Lab — Jul. 2019 - Jun. 2021
- Designed a data-informed task-agnostic attention pruning method for transformer models. Evaluated performance on various models (autoregressive, autoencoder, or seq-to-seq transformers) and application areas (language modelling, translation, natural language understanding, question answering). Used sparse GPU kernels to lower memory footprint by 30% and increase inference speed by 10%.
- Defined a few-shot multi-task conditional image generation benchmark by leveraging structure in a large-scale storm event dataset. Improved performance of conditional GANs using meta-learning algorithms and contrastive pretraining.

Research Intern at Celixir
Stratford-upon-Avon, UK — Jun. 2018 - Aug. 2018
- *Overview:* regenerative medicine founded by Nobel Laureate Professor Sir Martin Evans
- Performed cell image analysis (segmentation, feature extraction) to predict cell culture health.
- Estimated feature importance to design future experiments and reduce number of assays.
- Collaborated with biologists to incorporate expert priors for Bayesian inference with MCMC simulations.

Research Intern at Shell Technology Centre Bangalore
Bangalore, India — Jun. 2017 - Aug. 2017
- Skeletonized 3D voxel grids using either their distance transforms or thinning algorithms.
- Implemented and evaluated heuristics for graph search algorithms to speed up numerical simulations.

EDUCATION

Massachusetts Institute of Technology, Cambridge MA
MEng. in EECS (5.0/5.0 GPA) , B.S. in EECS and Physics — Sep. 2015 - May 2021
- *Selected CS coursework:* Algorithms, Machine Learning, Optimization for ML; Bayesian Modeling, Meta-Learning, Statistics Computation & Applications; Computer Systems, Software Construction.
- *Selected Physics coursework:* Quantum physics 1-3, General Relativity, Experimental Physics, Statistical Physics 1
- *Teaching Experience*: Teaching Assistant for Computation Structures (6.004).

ACADEMIC ACHIEVEMENTS

Peer Reviewed Publications:

- Adib Hasan, **Ileana Rugina**, Alex Wang; *Pruning for Protection: Increasing Jailbreak Resistance in Aligned LLMs Without Fine-Tuning* ; EMNLP 2024 Workshop BlackBoxNLP
- **Ileana Rugina**, Rumen Dangovski, Li Jing, Preslav Nakov, Marin Soljačić; *Data-Informed Global Sparseness in Attention Mechanisms for Deep Neural Networks*; LREC-Coling 2024
- **Ileana Rugina**, Rumen Dangovski, Mark Veillette, Pooya Khorrami, Brian Cheung, Olga Simek, Marin Soljačić; *Meta-Learning and Self-Supervised Pretraining for Storm Event Imagery Translation*; IEEE High Performance Extreme Computing Conference 2023; earlier version presented at ICLR AI for Earth and Space Science Workshop 2022
- Pooya Khorrami, Olga Simek, Brian Cheung, Mark Veillette, Rumen Dangovski, **Ileana Rugina**, Marin Soljačić, Pulkit Agrawal ; *Adapting Deep Learning Models to New Meteorological Contexts Using Transfer Learning*; IEEE International Conference on Big Data 2021

Silver Medal - Asian Physics Olympiad; **Bronze Medal** - International Physics Olympiad 2015

CLASS/PERSONAL PROJECTS

- SWE: implemented mapreduce, subset of raft consensus in go; networked multiplayer Pinball game in java.
- ML Theory: Last Iterate Convergence of EG Methods for Variationally Coherent Min-Max Problems.

SKILLS AND INTERESTS

- Python: ML (numpy, pytorch, jax, sklearn, pymc) and web (asyncio, tornado, FastAPI, Django)
- ML: NLP; speech recognition and synthesis; data-efficient (contrastive, few-shot) learning; bayesian learning
- Cloud & infra: GCP, AWS, docker, kubernetes, redis, rabbitmq, Prometheus, Grafana, terraform, nginx
- familiar programming languages: Go, JS and React, scss, SQL, bash

Interests: Algorithmic and Systems methods for efficient inference; Distributed Systems.