Shang Gao

Research Scientist • Biostatistics and Multiscale Systems Group
Oak Ridge National Laboratory • 1 Bethel Valley Road, Oak Ridge, TN, USA
Email: emailshang@gmail.com • Github: github.com/iamshang1/Projects
Last updated: November 17, 2021

Summary

I develop novel, scalable deep learning and natural language processing solutions for large scale, real-world applications. I am experienced and knowledgeable in a wide range of machine learning and data science techniques and can effectively apply these to address challenging, practical problems with various data, compute, and other limitations. I am currently the primary technical lead of a major research collaboration between the National Cancer Institute and U.S. Department of Energy on utilizing AI and HPC to automate cancer surveillance. My recent work includes developing a novel, state-of-the-art deep learning architecture for extracting key data elements from cancer pathology reports, scaling Transformer language models for pretraining and fine-tuning on ORNL's Summit supercomputer, and developing a transfer learning and semi-supervised approach for biomedical named entity recognition in low data settings.

Research Experience

2021–now **Scalable AI and NLP Project Lead**, Research Scientist, Biostatistics and Multiscale Systems Group, Oak Ridge National Laboratory

- Primary technical lead of 10+ member research team working on multi-year collaboration with National Cancer Institute to automate cancer surveillance using deep learning and scalable computing
- Developing novel algorithms for knowledge-oriented pretraining of Transformer language models on long clinical and biomedical text
- Developing capabilities for massively distributed pretraining of custom Transformer architectures on Summit and Frontier leadership-class supercomputers

2020–2021 **Scalable AI and NLP**, Postdoctoral Researcher, Multimodal Data Analytics Group, Oak Ridge National Laboratory

- Develop and deploy deep learning architectures for automated information extraction from cancer pathology reports; techniques include CNN, RNN, and Transformer-based approaches
- Develop autonomous driving agents in CARLA simulation environment using end-to-end deep imitation learning and deep reinforcement learning methods
- Develop and apply multi-task, transfer learning, semi-supervised, visualization/interpretability, and uncertainty quantification methods for deep learning models
- Scale deep learning algorithms across multiple GPUs and nodes on Oak Ridge supercomputer clusters

2017–2019 **Deep learning for clinical NLP**, Graduate Researcher, Bredesen Center, The University of Tennessee, Knoxville

- Developed new state-of-the-art text classification model for cancer pathology reports based on neural self-attention; the approach achieves better accuracy and trains over 10x faster than the previous state-of-the-art method
- Developed visualization tool for interpretable deep learning for clinical text classification using neural attention weights
- Developed novel methodology to identify and correct for mismatches between human expert annotations and the content reported within individual cancer pathology reports; this method improves classification accuracy on cancer pathology reports by up to 10%

2017 **Frameworks for scalable AI**, ORISE Higher Education Research Experiences Intern, Oak Ridge National Laboratory

- Designed and implemented Python API to manage scientific-scale dataset transfers using Globus and CKAN framework
- Implemented deep learning pipeline to automate parameter search and optimization for deep learning models on Oak Ridge supercomputer clusters

2016 **Deep learning for health applications**, Graduate Researcher, Institute for Artificial Intelligence, The University of Georgia

- Worked with interdisciplinary team on human activity recognition project to classify activity type based on hip-worn accelerometer device
- Developed convolutional-LSTM model that achieves competitive performance on human activity recognition tasks without requiring manual engineering of features

Professional Experience

2012–2016 **Technical documentation and online training development**, Technical Writer, Noble Systems

- Produce customer-facing online training for a wide range of contact center products, including campaign management software, interactive voice response scripting interfaces, and more
- Maintain and develop structure, templates, procedures, and single-sourcing guidelines for internal, value added resellers, and customer knowledge bases – content includes product technical specifications, client connectivity information, troubleshooting procedures, database reference tables, and best practices
- Work with Engineering, Development, and Support teams to produce troubleshooting and configuration guides for both internal and customer use
- Troubleshoot all technical issues related to internal and customer knowledge bases, including issues with HTML/CSS formatting, Team Foundation Server version control, and nightly auto-build and publishing process

Education

Doctor of Philosophy, Data Science and Engineering, Awarded 12/2019
The University of Tennessee Knoxville, Bredesen Center, Knoxville, TN

- Thesis topic: *Hierarchical Neural Architectures for Classifying Cancer Pathology Reports*

- Research interests: deep learning for clinical natural language processing
- Advisers: Dr. Arvind Ramanathan, Argonne National Laboratory & Dr. Georgia Tourassi, Oak Ridge National Laboratory

Bachelor of Science, Economics, Graduated 05/2009
Duke University, Durham, NC
- Minor: film, video, and digital production
- William J Griffith University Service Award for Outstanding Contributions to the Duke Community
- Distinguished Leadership and Service Award for Expanding the Boundaries of Learning
- Hal Kammerer Memorial Award for Outstanding Film and Video Production

Accomplishments and Awards

2021	**Gordon Bell Finalist for "Language Models for the Prediction of SARS-CoV-2 Inhibitors"**, Supercomputing 2021
2021	**INCITE Award Co-PI for "Scalable Transformer language models for drug discovery"**, Oak Ridge Leadership Computing Facility
2021	**ALCC Award Co-PI for "Next-generation scalable deep learning for medical natural language processing"**, Oak Ridge Leadership Computing Facility
2020	**ALCC Award Co-PI for "Evolutionary Multi-scenario Simulation Environment for Autonomous Vehicle Testing"**, Oak Ridge Leadership Computing Facility
2018	**Entrepreneurship Award**, Bredesen Center
2018	**Fall I-Corp South Regional Cohort Alumni**, National Science Foundation
2017	**Most Novel Solution**, Smoky Mountain Data Challenge

Teaching and Advising

2020-2021	**Co-Advisor:** Kevin De Angeli, Bredesen Center, currently a graduate student at the University of Tennessee, Knoxville
2021	**Co-Lecturer:** Intro to Data Science Graduate Level Course, The University of Tennessee, Knoxville
2018 - 2019	**Guest Lecturer:** Deep Learning Course, University of Tennessee, Knoxville

Community Outreach

2018-2019	Guest Lecturer for various seminars and courses at the Bredesen Center Data Science and Engineering program at the University of Tennessee, Knoxville
2018	Volunteer at "Traveling Science Fair" events at American Museum of Science and Energy, Oak Ridge, Tennessee
2017–2018	Volunteer at Oak Ridge Computer Science Girls classes, Oak Ridge, Tennessee
2017	DaVinci Art & Science Fair judge, Jefferson Middle School, Oak Ridge, Tennessee
2017	Volunteer at "Introduce Your Daughter to Code" Women in Computing @ ORNL event, Oak Ridge National Laboratory, Tennessee

Patents

2019	**Live Call Debugging and Monitoring Tool for an Interactive Voice Response Unit**, US10212283
2016	**Utilizing Predictive Models to Improve Predictive Dialer Pacing Capabilities**, US9723144B1

Publications

Journal Articles

Angeli, Kevin De, **Shang Gao**, Mohammed Alawad, Hong-Jun Yoon, Noah Schaefferkoetter, Xiao-Cheng Wu, et al. (2021). "Deep active learning for classifying cancer pathology reports." In: *BMC Bioinformatics* 22.1, pp. 113–113.

Blanchard, Andrew, Mayanka Chandra Shekar, **Shang Gao**, John Gounley, Isaac Lyngaas, Jens Glaser, et al. (2021). "Automating Genetic Algorithm Mutations for Molecules Using a Masked Language Model". In: *Under Review*.

Blanchard, Andrew, **Shang Gao**, Hong-Jun Yoon, Blair Christian, Eric Durbin, Xiao-Cheng Wu, et al. (2021). "A Keyword-Enhanced Approach to Handle Class Imbalance in Clinical Text Classification". In: *Under Review*.

De Angeli, Kevin, **Shang Gao**, Ioana Danciu, Eric Durbin, Xiao-Cheng Wu, Antoinette Stroup, et al. (2021). "Class Imbalance in Out-of-Distribution Datasets: Improving the Robustness of the TextCNN for the Classification of Rare Cancer Types". In: *Accepted for publication in Journal of Biomedical Informatics*.

Gao, **Shang**, Mohammed Alawad, Michael Todd Young, John Gounley, Noah Schaefferkoetter, Hong-Jun Yoon, et al. (2021). "Limitations of Transformers on Clinical Text Classification." In: *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1.

Gao, **Shang**, Olivera Kotevska, Alexandre Sorokine, and J Blair Christian (2021). "A pre-training and self-training approach for biomedical named entity recognition". In: *PLOS ONE* 16.2.

Alawad, Mohammed, **Shang Gao**, John X Qiu, Hong Jun Yoon, J Blair Christian, Lynne Penberthy, et al. (2020). "Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks." In: *Journal of the American Medical Informatics Association* 27.1, pp. 89–98.

Alawad, Mohammed, Hong-Jun Yoon, **Shang Gao**, Brent Mumphrey, Xiao-Cheng Wu, Eric B. Durbin, et al. (2020). "Privacy-Preserving Deep Learning NLP Models for Cancer Registries". In: *IEEE Transactions on Emerging Topics in Computing*, pp. 1–1.

Gao, **Shang**, Mohammed Alawad, Noah Schaefferkoetter, Lynne Penberthy, Xiao-Cheng Wu, Eric B. Durbin, et al. (2020). "Using case-level context to classify cancer pathology reports". In: *PLOS ONE* 15.5, pp. 1–21.

Yoon, Hong-Jun, Hilda B. Klasky, John P. Gounley, Mohammed Alawad, **Shang Gao**, Eric B. Durbin, et al. (2020). "Accelerated training of bootstrap aggregation-based deep information extraction systems from cancer pathology reports." In: *Journal of Biomedical Informatics* 110, p. 103564.

Gao, **Shang**, John X. Qiu, Mohammed Alawad, Jacob D. Hinkle, Noah Schaefferkoetter, Hong-Jun Yoon, et al. (2019). "Classifying cancer pathology reports with hierarchical self-attention networks." In: *Artificial Intelligence in Medicine* 101, p. 101726.

Bhowmik, Debsindhu, **Shang Gao**, Michael T. Young, and Arvind Ramanathan (2018). "Deep clustering of protein folding simulations." In: *BMC Bioinformatics* 19.18, pp. 47–58.

Gao, **Shang**, Michael T. Young, John X. Qiu, Hong-Jun Yoon, James Blair Christian, Paul A. Fearn, et al. (2018). "Hierarchical attention networks for information extraction from cancer pathology reports." In: *Journal of the American Medical Informatics Association* 25.3. Journal Article, pp. 321–330.

Conference Articles

Blanchard, Andrew, John Gounley, Debsindhu Bhowmik, Mayanka Chandra Shekar, Isaac Lyngaas, **Shang Gao**, et al. (2021). "Language Models for the Prediction of SARS-CoV-2 Inhibitors". In: *Supercomputing 2021*.

Gao, **Shang**, Spencer Paulissen, Mark Coletti, and Robert Patton (2021). "Quantitative Evaluation of Autonomous Driving in CARLA". In: *From Benchmarking Behavior Prediction to Socially Compatible Behavior Generation in Autonomous Driving Workshop at 32nd IEEE Intelligent Vehicles Symposium*.

Agrawal, Devanshu, **Shang Gao**, and Jacob Hinkle (2020). "Bayesian Deep Learning for Robust Information Extraction from Cancer Pathology Reports". In: *2020 Computational Approaches for Cancer Workshop at Supercomputing (CAFCW)*.

Alawad, Mohammed, **Shang Gao**, Folami Alamudun, Xiao-Cheng Wu, Eric B Durbin, Jennifer Doherty, et al. (2020). "Multimodal Data Representation with Deep Learning for Extracting Cancer Characteristics from Clinical Text". In: *2020 IEEE International Conference on Big Data*.

Alawad, Mohammed, **Shang Gao**, Mayanka Chandra Shekar, S M Shamimul Hasan, J Blair Christian, Georgia Tourassi, et al. (2020). "Integration of Domain Knowledge using Medical Knowledge Graph with Deep Learning for Cancer Phenotyping". In: *2020 Computational Approaches for Cancer Workshop at Supercomputing (CAFCW)*.

Patton, Robert, **Shang Gao**, Spencer Paulissen, Nicholas Haas, Brian Jewell, Xiangyu Zhang, et al. (2020). "Heterogeneous Machine Learning on High Performance Computing for End to End Driving of Autonomous Vehicles". In: *SAE Technical Paper Series*.

Alawad, Mohammed, **Shang Gao**, John Qiu, Noah Schaefferkoetter, Jacob D. Hinkle, Hong-Jun Yoon, et al. (2019). "Deep Transfer Learning Across Cancer Registries for Information Extraction from Pathology Reports". In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 1–4.

Alawad, Mohammed, **Shang Gao**, Xiao-Cheng Wu, Eric B. Durbin, Linda Coyle, Lynne Penberthy, et al. (2019). "Adversarial Training for Privacy-Preserving Deep Learning Model Distribution". In: *2019 IEEE International Conference on Big Data (Big Data)*, pp. 5705–5710.

Qiu, John X., **Shang Gao**, Mohammed Alawad, Noah Schaefferkoetter, Folami Alamudun, Hong-Jun Yoon, et al. (2019). "Semi-Supervised Information Extraction for Cancer Pathology Reports". In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 1–4.

Yoon, Hong-Jun, John Gounley, **Shang Gao**, Mohammed Alawad, Arvind Ramanathan, and Georgia Tourassi (2019). "Model-based Hyperparameter Optimization of Convolutional Neural Networks for Information Extraction from Cancer Pathology Reports on HPC". In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 1–4.

Gao, **Shang**, Arvind Ramanathan, and Georgia D. Tourassi (2018). "Hierarchical Convolutional Attention Networks for Text Classification". In: *Proceedings of The Third Workshop on Representation Learning for NLP*, pp. 11–23.