Logan Riggs Smith

logansmith5@gmail.com

Experience

Sept. 2022 – Present: Independent Researcher – Working on various projects with the overall aim of automating alignment research and interpreting language models. Funded by Long Term Future Fund.

- Visualizing intermediate layer logits of language models with learned unembeddings to understand memorization and prompt injection
- Formatting the alignment dataset for finetuning language models to learn alignment research terms, comment structure, and mimic specific researcher's forum interaction
- Overseeing two research fellows in mechanistic interpretability in language models

Jan. 2022 – July 2022: Independent Researcher – Collected and published a contemporary dataset of AI Alignment literature with 89M tokens. Specifically, created and tested a scraper for the forum lesswrong.com which captures relevant meta-data, recursive-comment structure, and LaTeX. Funded by Long Term Future Fund.

Jan. 2021 – May 2021: Participant in AI Safety Camp – Proposed and led a research agenda on making neural networks modular to improve interpretability. Coordinated logistics and minutes for meetings for our team of six and presented findings.

May. 2019 – May 2021: Graduate Research Assistant, Mississippi State University – Coordinated with a group of 4-6 researchers, presenting updates on a weekly basis. Suggested and implemented machine learning experiments using TensorFlow. Researched, wrote, and published technical papers.

- Wrote literature reviews for wireless physical fingerprints, zero-shot learning, and outlier detection
- Trained custom machine learning networks including MLPs, CNNs, and auto-encoders
- Pre-processed wireless signals with normalization, the Fourier transform, and the short-time Fourier transform

Jan. 2019 – May. 2019: Teaching Assistant, Mississippi State University – Assisted undergraduates in their Microprocessor lab. Submitted grades within a few days of student submission. Personally completed the lab assignment the week before to better assist students.

Nov. 2017 – Aug. 2018: Software Developer (co-founder), WISPr Systems – Researched and implemented software solutions to allow drones to capture internet signal strength, displaying the results with React. Created simulated autonomous flights plans and tested in real-world situations.

Publications

Kirchner, J. H., Smith, L., Thibodeau, J., McDonnell, K., and Reynolds, L. *Understanding AI alignment research: A Systematic Analysis.* arXiv preprint arXiv:2206.02841 (2022).

Turner, A. M., Smith, L., Shah, R., Critch, A., and Tadepalli, P. *Optimal Policies Tend to Seek Power* in Advances in Neural Information Processing Systems 34 pre-proceedings (NeurIPS 2021) (Spotlight Paper)

Smith, L., Smith, N., Kodipaka, S., Dahal, A., Tang, B., Ball, J. E., and Young, M. *Effect of the Short Time Fourier Transform on the Classification of Complex-Valued Mobile Signals* in [Signal Processing, Sensor/Information Fusion, and Target Recognition XXX], 11756, International Society for Optics and Photonics (2021)

Smith, N., Smith, L., Kodipaka, S., Dahal, A., Tang, B., Ball, J. E., and Young, M. *Real-Time Location Fingerprinting for Mobile Devices in an Indoor Prison Setting* in [Signal Processing, Sensor/Information Fusion, and Target Recognition XXX], 11756, International Society for Optics and Photonics (2021)

Smith, L., Smith, N., Rayborn, D., Tang, B., Ball, J. E., and Young, M. *Identifying unlabeled wifi devices with zero-shot learning* in [Automatic Target Recognition XXX], 11394, 113940R, International Society for Optics and Photonics (2020).

Smith, L., Smith, N., Hopkins, J., Rayborn, D., Ball, J. E., Tang, B., and Young, M.*Classifying wifi "physical fingerprints" using complex deep learning* in [Automatic Target Recognition XXX], 11394, 113940J, International Society for Optics and Photonics (2020).

Oral Presentations

Identifying unlabeled wifi devices with zero-shot learning SPIE defense + Commercial Sensing Digital Forum (24 April 2020)

Effect of the Short Time Fourier Transform on the Classification of Complex-Valued Mobile Signals SPIE defense + Commercial Sensing Digital Forum

Education

MS Computer Engineering, Mississippi State University, 2021, Final Grade: 3.85/4.0

BSc Computer Engineering, Mississippi State University, 2018. Final Grade: 3.68/4.0