Joshua L. Shapiro

(202) 351-1972 | joshua.shapiro@me.com | jshapiro.info

Experience

Staff Machine Learning Engineer November 2024 - Present
AKASA Washington, DC

- Leading the ML development for AKASA's medical coding solution, accelerating the research-to-production timeline.

Senior Machine Learning Engineer January 2022 - October 2024
AKASA Washington, DC

- Developed a compute-agnostic job launcher and internal machine learning training framework to standardize distributed model training across environments. Introduced Weights & Biases for experiment tracking.
- Established standardized offline ML evaluation framework for the medical coding solution, enabling direct comparison of offline experiments to production results, increasing research iteration speed.
- Standardized inference artifact interface and implementation for the medical coding solution, reducing iteration lifecycle from weeks to days.
- Developed ML inference engine for the medical coding solution, capable of serving LLMs upwards of 10B parameters. Reduced latency by 7x by leveraging VLLM.

Industry Faculty August 2023 - Present
The George Washington University, School of Engineering & Applied Science Washington, DC

- Teaching the Senior Design Capstone course for the Computer Science department, guiding students through end-to-end software development projects while emphasizing industry best practices.

Lead Research Engineer April 2020 - December 2021
ASAPP New York, NY

- Led initiative to generate rich coversational embeddings, increasing performance across multiple production models.
- Researched novel attention-based RNN architecture for hybrid ASR and applied model to production speech to text services.
- Implemented quickthought-style RNN training regime that decreased model size while increasing performance across a variety of production classification tasks.
- Co-developed internal machine learning training framework to speed up the research to production pipeline.

Senior Machine Learning Engineer July 2018 - March 2020
ASAPP New York, NY

- Designed and implemented entity recognition, conversation summarization, and mid-flow branch classification services for dialogue systems. Collaborated with Research to productionize model prototypes, Product to define service requirements, and Data Science to run A/B tests.

Cognitive Software Engineer September 2017 - June 2018
IBM Research Yorktown Heights, NY

- Worked in the Data Centric Systems Department at the intersection of high performance computing and deep learning.
- Researched novel techniques for highly scalable video action classification that at the time outperformed state-of-the-art models in terms of accuracy and training speed.
- Created novel temporal state detection and clustering algorithms for molecular dynamics simulations.

Publications

ASAPP-ASR: Multistream CNN and Self-Attentive SRU for SOTA Speech Recognition 2020
J. Pan, J. Shapiro, J. Wohlwend, KJ. Han, T. Lei, T. Ma Interspeech

Video Action Recognition with an Additional End-to-End Trained Temporal Stream 2019
G. Cong, G. Domeniconi, J. Shapiro, CC. Yang, B. Chen IEEE WACV

Accelerating Deep Neural Network Training for Action Recognition on a Cluster of GPUs 2018
G. Cong, G. Domeniconi, J. Shapiro, F. Zhou, B. Chen SBAC-PAD

Education

The George Washington University 2013-2017
Bachelor of Science in Computer Science; GPA: 3.89 Washington, DC

Korea University Spring 2015
Exchange Program Seoul, South Korea

Technical Skills

Deep Learning: PyTorch, distributed training, Transformers, RNNs, Huggingface, Pytorch Lightning, NLP, LLMs

Programming: Python, Jupyter, SQL

Technical Tools: AWS, Docker, Kubernetes, Prefect, Jira, agile programming methodologies