

Joshua L. Shapiro

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

EXPERIENCE

Staff Machine Learning Engineer

AKASA

November 2024 - Present

Washington, DC

- Technical lead for the ML development of AKASA's medical coding solution, focusing on accelerating the research-to-production timeline.

Senior Machine Learning Engineer

AKASA

January 2022 - October 2024

Washington, DC

- Architected a compute-agnostic job launcher and internal machine learning training framework to standardize distributed model training across environments and projects. Introduced Weights & Biases for experiment tracking.
- Implemented a comprehensive offline ML evaluation framework for the medical coding solution, enabling direct comparison of offline experiments to production results, increasing research iteration speed.
- Streamlined inference artifact 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

The George Washington University, School of Engineering & Applied Science

August 2023 - Present

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

ASAPP

April 2020 - December 2021

New York, NY

- Led initiative to generate rich conversational 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

ASAPP

July 2018 - March 2020

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

IBM Research

September 2017 - June 2018

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

J. Pan, J. Shapiro, J. Wohlwend, K.J. Han, T. Lei, T. Ma

2020

Interspeech

Video Action Recognition with an Additional End-to-End Trained Temporal Stream

G. Cong, G. Domeniconi, J. Shapiro, CC. Yang, B. Chen

2019

IEEE WACV

Accelerating Deep Neural Network Training for Action Recognition on a Cluster of GPUs

G. Cong, G. Domeniconi, J. Shapiro, F. Zhou, B. Chen

2018

SBAC-PAD

EDUCATION

The George Washington University

Bachelor of Science in Computer Science; GPA: 3.89

2013-2017

Washington, DC

Korea University

Exchange Program

Spring 2015

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