Sam Anzaroot

☎ (718) 207-2887 ✉ samanzaroot@gmail.com
samanz.com samanz sam-anzaroot
Sam Anzaroot

WORK EXPERIENCE

Verneek

Feb. 2021 - Present | *Applied AI Researcher*

Founding engineer, first employee of the company. Built and launched the company's first product.

Built core machine learning/AI/NLP technologies using state-of-the-art transformer language models, such as Llama, Mistral and other open source LLM, as well as LLM apis such as ChatGPT, Claude, etc.

Created automatic training, evaluation, and deployment pipelines using Kubeflow and NVIDIA Triton Inference Server.

Implemented a comprehensive chat workflow system, for a deployed RAG inspired chat system for ecommerce that can perform product, informational, and customer service functional workflows.

Trained internal semantic parsing and code generation models with novel data augmentation methods to achive over 98% accuracy on internal datasets.

Researched and trained retrieval embedding models using unsupervised methods to significantly increase retrieval bi-encoder model performance for RAG and search applications.

Sped up large language models using model distillation and quantization for 10x speedup in inference time and throughput.

Helped managed AI team employees and mentored interns.

Dataminr

July 2019 - Sept. 2020	*Principal Data Scientist*
Jan. 2017	*Senior Data Scientist*
Feb. 2015	*Data Scientist*
Sept. 2014	*Software Engineer in Data*

Helped grow the AI team over six years by leading multiple high-profile projects, advocating internally for state-of-the-art techniques, leading an AI reading group and mentoring interns. Communicated with stakeholders including product managers, HCI researchers, designers, domain experts, and engineers.

Led and contributed to team focused on automatically generating summaries of public safety events detected from social media posts. The team utilized **seq2seq LSTM** and **Transformer** deep-learning models, and ran a user study and deployed a **human-in-the-loop** system for summary writing to production which sped up summary writing by 2x.

Led and contributed to geo-prediction team, focused on detecting mentions of locations in unstructured text and geocoding mentions to points on earth. Trained and deployed a **neural network conditional random field** model and **neural network LambdaRank** model, drastically increasing location precision on Dataminr content.

Led and contributed to automation team, combining multiple different models in a pipeline for full content automation. This project resulted in the full automation of the majority of content sent by Dataminr.

Worked as IC on various projects, including a novel language-identification model for social media, a text-based topic prediction model, a novel neural-network library built in Scala, a named entity recognition model for social media, and a label annotation platform.

Oracle Labs - East

Feb. 2014 - June 2014 | *Research Intern*

Researched methods for highly parallel probabilistic inference on **conditional random fields (CRFs)** using GPUs.

EDUCATION

University of Massachusetts — Amherst

Feb. 2014 | *MS in Computer Science*

Queens College — City College of New York

June 2011 | *BS in Computer Science*
Magna cum laude

Created a GPU version of the **belief propagation** algorithm written in **CUDA**. Optimized this implementation to allow for 200x speedup in inference and 100x speedup in training of **CRFs** over sequential implementation.

IESL, University of Massachusetts — Amherst

Sept. 2011 - Feb. 2014 | *Research Assistant*

Advisor: Andrew McCallum

Performed **NLP** and **ML** research focusing on **undirected graphical models** and **information extraction**.

Oversaw creation of a novel citation extraction dataset, the largest and most fine-grained openly available dataset for this task.

Developed method for more robust inference in conditional random fields using extensions to **Lagrange relaxation** methods called **soft dual-decomposition** with applications in citation extraction, retrieving new state-of-the-art results on the citation extraction task.

VOLUNTEER EXPERIENCE

Datakind

March 2016 - Sept. 2016 | *Data Science Volunteer*

Implemented methods for automatically extracting metadata from research documents to assist researchers in performing systematic literature reviews.

Helped build and deploy a machine learning enabled systematic review web application currently in use by researchers available at colandrapp.com

PUBLICATIONS

Chidubem Arachie, Manas Gaur, **Sam Anzaroot**, William Groves, Ke Zhang, and Alejandro Jaimes.
Unsupervised Detection of Sub-Events in Large Scale Disasters.
AAAI Conference on Artificial Intelligence, 2020.

Shan Jiang, William Groves, **Sam Anzaroot**, and Alejandro (Alex) Jaimes.
Crisis Sub-Events on Social Media: A Case Study of Wildfires
AI for Social Good Workshop at the 36th International Conference on Machine Learning (AISG@ICML 2019), 2019.

Cheng SJ, Augustin C, Bethel A, Gill D, **Anzaroot S**, Brun J, DeWilde B, Minnich RC, Garside R, Masuda YJ, Miller DC, Wilkie D, Wongbusarakum S, McKinnon MC.
Using machine learning to advance synthesis and use of conservation and environmental evidence.
Conservation Biology, 2018.

Sam Anzaroot, Alexandre Passos, David Belanger, Andrew McCallum.
Learning Soft Linear Constraints with Application to Citation Field Extraction.
52nd Annual Meeting of the Association for Computational Linguistics (ACL2014), 2014.

Sam Anzaroot, Andrew McCallum.
A New Dataset for Fine-Grained Citation Field Extraction
ICML Workshop on Peer Reviewing and Publishing Models (PEER), 2013.

Qi Li, **Sam Anzaroot**, Wen-Pin Lin, Xiang Li and Heng Ji.
Joint Inference for Crossdocument Information Extraction
20th ACM Conference on Information and Knowledge Management (CIKM2011), 2011.

tome

SKILLS

Programming Languages

- Scala
- Java
- Python
- Javascript
- C++
- C
- CUDA
- Bash

Machine Learning Frameworks

- PyTorch
- Tensorflow

Data Processing

- Spark
- Hadoop
- Postgresl - SQL

Tools

- Node
- JQuery
- JIRA
- Git
- Docker
- LaTeX

Data Science

- Machine learning
- Deep learning
- Natural language processing
- Ranking