Ashutosh Tiwari

669-292-7534 | findashutoshtiwari@gmail.com | Linkedin@ashutosh–tiwari | Homepage/Portfolio | Github@thunderock

Work Experience

Machine Learning Engineer 4 (ML Inference Platform) Jul. 2024 – Present
ADOBE FIREFLY *ML PLATFORM, GENERATIVE AI, LLMS*

- Building ML inference platform for offline batch inference serving generative AI workloads at scale for videos and images.
- Architected inference framework using PyTorch Lightning for distributed model serving across GPU clusters.
- Leading development of inference framework leveraging vLLM, Ray Serve, and Ray Data for distributed inference of large language models (LLMs), enabling efficient serving of multi-billion parameter models.
- Integrated TensorRT-LLM and Triton Inference Server for optimized model serving with dynamic batching and GPU optimization.
- Designed feature store infrastructure supporting training of ML models on billions of images and videos.

Senior Software Engineer (Data Platform) Sep. 2023 – Jul. 2024
EVOLUTIONIQ *DATA PLATFORM, NATURAL LANGUAGE PROCESSING, REGRESSION*

- Built data pipelines for ML model training at intersection of generative AI, fintech, and healthcare.
- Developed pipelines to train models for claim duration prediction, ICD extractions, and VOC recommendations.
- Created fairness evaluation framework ensuring models are unbiased across demographic groups.

Software Dev Engineer II (ML Platform) Jan. 2019 – Jul. 2021
SWIGGY *ML PLATFORM, TIME SERIES FORECASTING, NLP*

- Led Data Acquisition Platform (DAQ) for capturing and scheduling APIs at scale, processing 15M rows daily for ML model training and analysis. Architecture included proxy service, request generation, transformation, and scraping modules using Scrapyd.
- Built Feature Store pipeline feeding on-demand features to deployed ML models at production scale (4Bn rows, 10K QPS) with multichannel ingestion via Spark, Flink, and file uploads.
- Founding member of Forecasting and Correlation Platform using Facebook Prophet and Polynomial Regression, powering critical scaling decisions with real-time time series forecasts across organization.
- Applied advanced NLP techniques for automated food tagging in customer-agent chats as part of AI democratization program at Swiggy, improving service efficiency.

Software Development Engineer (Search Relevance) Sep. 2017 – Jan. 2019
FLIPKART (Walmart) *Search Relevance, ML Ops, NLP*

- Built and improved CRF and Neural Network-based search intent models in production, powering search and discovery for millions daily. Identified error classes, developed solutions with Data Scientists, and deployed fixes.
- Implemented FastText-based query store classifier predicting category for tail queries, later replaced with bi-list based model.
- Created first automated ML training and deployment workflow in Flipkart using Luigi/Airflow, orchestrating end-to-end flow with data and model validations. Built generic framework generating dynamic DAGs for different ML models at runtime.
- Developed Flask APIs to serve ML intent models in production, ensuring low-latency inference for search queries.
- Built large-scale pipelines (4Bn+ datapoints) using Cascading/HDFS for extracting and transforming user events into training data.
- Winner of Hackday 9 (Flipkart-wide Hackathon 2018) in Ekart category - built multi-label image annotation model inspired by Inception v3 for lifestyle products, achieving 99.6% validation accuracy using differential learning rates.

Software Development Engineer Sep. 2016 – Sep. 2017
GROUPON *Backend Engineering, Microservices*

- Developed Cyclops, customer service interface used by representatives globally to resolve queries and requests across all Groupon countries.
- Built integrations with multiple microservices and exposed REST APIs to internal services using Ruby on Rails, CoffeeScript, and MySQL.

Software Engineer Sep. 2015 – Aug. 2016
NETSPEED SYSTEMS (Intel) *Graph Algorithms, Network on Chip*

- Implemented critical SOC simulation modules single-handedly: Virtual Channel Arbitration, Polarity-based Arbitration, Multi-Cast Filtering, and Structural Latency Breakdown for AXI and Streaming protocols using C++.

Publications

Accepted at *NetSci 2023* (Poster Presentation) and *IC2S2 2023* (Parallel Talk) as **first author**

[1] **Ashutosh Tiwari**, Prof. Sadamori Kojaku, Prof. Yong-Yeol Ahn, "Biased Contrastive Learning debiases Graph Neural Networks," *International Conference on Network Science (NetSci)*, 2023. Poster

In this work, we propose a non-parametric contrastive learning framework to learn debiased graph embeddings with respect to sensitive node attributes and structural homophily. Through empirical evaluations on different datasets, we demonstrate that our method offers a better approach to debiasing compared to existing approaches and thus results in more organic recommendations across different GNN architectures.

Education

Indiana University, Bloomington Bloomington, IN
Master of Science in Computational Data Science 3.87/4.0

National Institute of Technology Patna, India
Bachelor of Science in Computer Science 8.32/10.0

Selected Projects

Graph ML | *PyTorch, PyTorch Geometric, Graph Neural Networks (GNN), Machine Learning on Graphs* Code

- Sklearn-style API for machine learning on graphs using PyTorch and PyTorch Geometric. Supports graph embeddings, graph sampling, and graph neural networks for recommendation systems.

BiasNet | *Deep Reinforcement Learning (RL), PyTorch, Actor Critic Algorithm, On-policy Model Free* Code/Report

- Learning to fight in Street Fighter II with induced relational bias from differential game scenes.

Investigating Bias Manifolds | *Bias Manifolds, Bias Progression, Measuring Bias, Python, Word2vec* Code/Report

Continuous Dominant Set Repair | *C++, Graph Algorithms, Guha and Khuller's Algo., Greedy* Code

- Repairs a broken link in Continuous Dominant Set in $O(\Delta^2)$, where Δ being the avg cardinality of connected graph.

BlindNet | *Python, PyTorch, Deep Learning, Transfer Learning, Computer Vision* Code/Report

- Image to vector generation on Coco dataset using deep learning and transfer learning techniques.

DeepFoodie | *Python, TensorFlow, Self Supervised Deep Clustering, Deep Learning, Transfer Learning* Code/Report

- Clustering dishes on basis of their ingredient embeddings generated by a neural network using self-supervised deep clustering.

Technical Skills

ML Inference & Serving: vLLM, Ray Serve, Ray Data, TensorRT-LLM, Triton Inference Server, Model Optimization, Distributed Inference, Batch Inference, GPU Optimization, Dynamic Batching
Machine Learning: Large Language Models (LLMs), Graph Neural Networks (GNN), Natural Language Processing (NLP), Computer Vision (CV), Deep Learning (DL), Reinforcement Learning (RL), Feature Engineering, Model Training, Fairness Aware ML
Frameworks/Libraries: PyTorch, PyTorch Lightning, PyTorch Geometric, TensorFlow, Scikit-learn, Keras, vLLM, Ray, Numpy, Pandas, Matplotlib
Data Engineering: Apache Spark, Apache Kafka, Apache Flink, Hadoop HDFS, Feature Store, Data Pipelines, ETL, Airflow, Luigi
Programming Languages: Python, Scala, C++, SQL, Java
Cloud & Infrastructure: AWS SageMaker, AWS DynamoDB, GPU Clusters, Distributed Systems, Docker, Kubernetes
Other Tools: Faiss, Cerberus, Django, Git, REST APIs, FastText, CRF, Cascading