Senthilkumar Gopal

sengopal.me | senthil777@gmail.com | linkedin.com/in/senthilkumargopal | 408-839-3815

Experience

Senior Applied Science Manager - ML Inference, AWS - Cupertino, USA Oct 2023 - Present

- Leading Neuron Science team for developing novel Inference acceleration methods on AWS accelerators for PyTorch.
- Managing applied researchers for improving LLMs inference performance using techniques such as weight/cache quantizations, speculative decoding techniques, continuous batching, flash attention etc.,
- Delivered critical projects for accelerating both internal (Bedrock, Alexa) and open source LLMs (Llama, Mistral, GPT etc.,) for better than GPU based price performance.
- Building benchmarks for ongoing evaluation of LLM performance, scalability, perplexity and divergence quality based on Open LLM Leaderboard.
- Developing distributed inference using novel sharding and parallelism techniques to support large scale multi-core and multi-instance deployments.
- Delivered HuggingFace equivalent *transformers-neuronx* focusing on generative large language models such as Llama, Mistral, GPT etc., saving inference costs over comparable GPU-based instances.
- Upstreaming core capabilities and integrations to vLLM, PyTorch XLA and other open source contributions.

Senior Applied Science Manager - Search Science, eBay Inc - San Jose, USA June 2021 – Sept 2023

- Managing applied researchers and machine learning engineers driving critical projects for eBay Search focusing on Visual Shopping, improving retrieval and building foundational embedding models
- Delivered XXM iGMB lift using token and embedding based extraction techniques by reducing null-low queries and improving recall relevance
- Delivered query-item similarity based ranking features leading to XXXM iGMB and built ML ranking features based on item characteristics and quality factors for title, aspects and image quality
- Developing Aspect quality models (seq2seq) to better qualify aspects to drive organized search, navigation, deterministic sorts, and inventory discovery
- Delivered image-based search using computer vision-based image embedding models (ResNet50) leading to 17% increase in DAU and developing integrated capabilities to improve relevance, discovery, and sorting functionalities
- Implementing embedding-based retrieval with sparse methods (bm25) to bring in semantically matching high-quality items, resulting in a 13% improved conversion for select verticals
- Responsible for building batched and near real-time data pipelines (Spark, Kafka) processing (*3B+ events per day*) to deliver search retrieval features such as aspect quality, image quality, modality alignment.
- Serving as the engineering leader to drive discovery for new visual shopping experiences, powered by image and multi-modal embeddings, object detection, and semantic cluster harvesting

Senior Applied Science Manager - Promotions Platform, eBay Inc - San Jose, USA May 2020 – June 2021

- Built a new team of applied researchers, technical leads and individual contributors to deliver high quality critical initiatives such as Coded coupons (600M in iGMV) and other promotion capabilities in Seller Marketing Platform
- Delivered XGBoost based models for ranking of promotional offers, for accurate painting in search results to improve promotion efficiency and drive purchase conversion
- Built end to end pipelines for daily promotion rescoring using Hadoop/Spark and obsolete promotion intervention
- Developed a multi-class classifier based recommendation system to drive new promotion generation leading to 20% increase in promotions using seller inventory and past promotions historical data.
- Built user embeddings (*item titles based context*) and clustering for creating buyer groups based on sales and search data, for targeted promotion campaigns leading to 4% increase in promotions usage.

Senior Applied Science Manager - Seller Analytics, eBay Inc - Toronto, Canada Jan 2019 – Apr 2020

- Built a team of data scientists, ML and full stack engineers, contributing to the seller analytics data platform and generating incremental revenue of over 100M in a year

- Delivered competitor analysis features using item and sales similarity embedding models increasing MAU by 5%
- Responsible for big data delivery pipelines (4B+ events/day) for near real time (Kafka) and batch loading transactional information (Spark) to power data analytics for sellers using ElasticSearch
- Implemented a dynamic pricing engine that enables sellers to adjust prices automatically and create new listings in near real-time capitalizing on early seasonal demands (*patent pending*)
- Developed LSTM and NER (NLP) based item title creation system with keyword suggestions to build relevant and high-performing keywords for product listings leading to 10% increase in subscriptions

Principal Identity Architect, eBay Inc - San Jose, USA Nov 2014 – Dec 2018

- Led the transformation of Identity Architecture, focusing on external Identity provider integrations, leveraging ML for bot mitigation, account takeover prediction and improved user experiences
- Lead Architect for Authentication Platform products consisting of Federated Identity, OAuth systems, Secure Token Services and Session Management which secure eBay internal/public APIs and Web Security
- Implemented multi-level caching Architecture and distributed computing capabilities to optimize performance/ UX

Member of Technical Staff, (Consultant), eBay Inc - Bangalore, India Jan 2010 – Nov 2014

- Built async microservices architecture for vendor communication and profile evaluation use cases
- Developed architecture for image upload, security protocols for vendor integration & storage for ID verification
- Developed a new Orchestration framework which produces dynamic verification flows with minimal configuration

Senior Software Engineer, Mphasis FinSolutions Pvt Ltd / Infosys Technologies Ltd - Chennai, India Nov 2005 – Jan 2010

- Developer Lead for building backend capabilities for dynamic generation of web applications (Google Maps Navigation, Calendar integration) in the Content Management System serving more than 1B page visits
- Developed job scheduling systems, PDF generation and backend REST/SOAP APIs with database integrations for full stack banking loan processing systems serving more than 100M users per day

Education

Coursera / DeepLearning.AI

Machine Learning, Deep Learning Specialization and NLP Specialization

Georgia Institute of Technology, USA

Master of Science, Computer Science, Machine Learning Specialization GPA: 4.0

Sathyabama Engineering College, Chennai India

Bachelor of Engineering, Electrical and Electronics GPA: 4.0

Awards

- Winner of Innovation Awards for “Generated Content for Seller Promotions” (2021), eBay Groups mobile app (2015), Vendor Integration Gateway (2012) and social data based application for eBay (2011)
- Winner of Exceptional Critical Talent Awards (awarded to less than 1%) at eBay for Seller Research Analytics (2019), Identity Platform (2018), Secure Token system (2016), Verification platform (2014), and global password reset (2014)

Publications

- Conference speaker on Scaling embedding models (MLOPsWorld, AIDevworld) and services authentication (IBM IndexConf, API World, Silicon Valley Code Camp, PRDC Deliver)
- Publications on eBay Applied Science on visual discovery with embeddings and others
- Open Source Contributor for eBay

Patents

- Deep navigation via a multimodal vector model - US Patent (*pending*)
- Dynamic pricing engine for early seasonal demands - US Patent (*pending*)
- Visual Search Query Intent Extraction and Search Refinement - US Patent (*pending*)