Avnish Kumar

+91 7233015412 | avnish.kumar.24.ak.ak@gmail.com | Linkedin | Github | Twitter | Blog

Education

Indian Institute of Technology (BHU) Varansi, India
B.Tech. in Chemical Engineering *Jun 2015 – May 2019*

Skills

Programming: Rust, C++, Python, JavaScript, LaTeX, AppleScript

Technologies: Git, Linux, Jenkins, Docker, Kubernetes, Airflow, Spark, Flink, MLFlow

Store & Index: Kafka, MongoDB, Redis, ElasticSearch, ELK, Scylla

Libraries: Pytorch (+Distributed), Hugginface, Fairseq, CVXOPT, Scikit-Learn, Tensorflow, Keras, Asyncio

Frameworks: FastAPI, Axum, Express, gRPC

Work Experience

Wynk Music Gurgaon, India
Machine Learning Engineer – RecSys, Music Representation Learning *Aug 2022 – Present*

- Developed a **Music Representation Learning** system using multi modal contrastive learning employing CLIP on BERT and Conformer Encoders with Pytorch DDP. Being used for music2music and music2tag retrieval. 7M+ acoustic representations indexed in ES on hybrid HSNW index. Label augmentation using LLMs for Indian Music.
- Developed a **Topic Recommendation system** for Music using indexed acoustic representation and Density Aware Locality Sensitive Hashing. Implemented highly scalable services in Rust to serve over **60M+ MAU** on Wynk.
- Developed a **Online Content Re-ranking system** for music recommendation products, boosting CTR by **20%**. Experimented with various point, pair, listwise approaches (e.g., DLRM, Mostra, xDeepFM, CatBoost, PNN, DCN, IFM).
- Implemented an Intent sampling based method for personalized recommendations behind the **"Surprise Me"** Feature on Wynk that servers 2500rps. Also servers as initializer for manu recommendation products and re-engagement campaigns.

Reliance Jio AI-CoE Hyderabad, India
Machine Learning Engineer – Speech, Anomaly-Detection, QP-Optimization *Jul 2019 – Aug 2022*

- JioMeet - Developed a **Real-time Speech-to-Text** system for Indian-English, Hindi, Telugu, Tamil, Marathi, and Bengali by finetuning Wav2Vec2.0 with domain-specific language models (KenLM) to achieve 7-8% WER across languages/domains. Experimented with incremental inference flows to reduce latency with minimal WER drop.
- Jio5G - Used telecom big-data to develop a scalable Optimiser for **Antenna-tilt automation** of 80,000+ cell towers in Mumbai, West-Bengal.A 10%+ better throughput and improved user-coverage.[**patent applied**]
- Designed and Developed a **Distributed Streaming Anomaly Detection** Engine used across multiple teams, For Cyber-Security it monitors 100,000+ servers for suspicious activity.

Algonomy Bangalore, India
Machine Learning Intern – Demand-Forecasting, Anomaly Detection, Look-Alike Modelling *May 2018 – July 2018*

- Worked with R&D team to develop a look alike model for marketing analytics in retail domain. Experiment with various distance metrics eg Mahalanobis, Euclidean, Cosine, etc. on a number of raw & engineered user attributes.
- Experimented with DeepAR and LSTM based models for demand forecasting and Random Cut Forest for Anomaly Detection on AWS Sagemaker.

Projects

Density Aware Locality Sensitive Hashing | Pytorch | *Paper* Feb 2024
Implemented Density aware Random Projections for Locality Sensitive Hashing in pytorch resulting in more homogeneous Hash Densities. These hashes can be used for cheap and swift filtering in retrieval tasks.

Obsidian Writing Assistant | TypeScript, Obsidian-APIs Dec 2023
A callout based plugin for obsidian that uses local LLMs(or APIs) to operate on enclosed text. Allows custom prompts to be passed along with selected text along with a number of presets available.

Chrome Plugin for Tab Indexing | JavaScript, Chrome-APIs, SQLite, FastAPI June 2023
Developed a Chrome plugin that allows users to manage their tabs, searching for tabs, and saving and indexing them for countless use-cases. User can make new tabs open on the left therefore all recent tabs are on the left and cmd+1/2/.., etc. can be used to switch between recent tabs. **Arxiv papers tab and downloaded file titles are renamed with paper title** so that they are easily searchable.

Token Merging for AST | Pytorch, ffmpeg, librosa | *Paper* April 2023
Experimented and Evaluated Token Merging on ASTs for inference time speedup. Findings show that token merging can reduce the number of tokens but reduces the accuracies significantly unlike segmentation models.

Edited: April 8, 2024