MRUGANK MILIND AKARTE

New York City (917) 900-8603
https://www.linkedin.com/in/mrugank-akarte/ mrugank.akarte@columbia.edu

PROFESSIONAL SUMMARY

Senior Machine Learning Engineer with 4+ years of experience bridging cutting-edge research and production-scale systems. Specialized in **distributed deep learning, graph neural networks, and large-scale recommendation systems**. Proven track record of adapting research techniques to real-world constraints, achieving training speedup through distributed optimization, and deploying novel architectures serving millions of users. Experience spans weak supervision learning, multi-level representation systems, and high-performance ML infrastructure.

Core Research Interests: Distributed ML Systems • Graph Neural Networks • Representation Learning • MLOps at Scale

TECHNICAL EXPERTISE

Deep Learning & Research:

• PyTorch • Tensorflow • Graph Neural Networks • Model Optimization

Distributed Systems:

• Ray (Multi-GPU/Multi-Node) • Kubernetes • Docker • Distributed Training • Data Parallelism • Pipeline Optimization

ML Infrastructure:

• Google Cloud Platform • Vertex AI • Kubeflow • TorchServe • MLOps • Model Serving • Real-time Inference

Programming & Data:

• Python • SQL • C++

RESEARCH & ENGINEERING EXPERIENCE

Macy's Technology New York City, USA
Senior Machine Learning Engineer- Product Recommendation Systems Feb 2021-Present

Graph Neural Networks for Cross-Category Recommendations - Researched and implemented GNN-based “Complete the Look” system using a novel weak supervision approach, achieving 3-4% improvement in revenue per visit and average order value across furniture catalog

- **Research Methodology:** Conducted literature review of state-of-the-art recommendation systems, adapting Amazon’s P-Companion architecture for cross-category compatibility prediction
- **Novel Training Strategy:** Designed innovative weak supervision framework combining collections metadata and co-purchase behavioral signals to address the absence of labeled compatibility data
- **Multi-level Architecture:** Built a hierarchical embedding system projecting product representations into category-aware compatibility space, enabling structured reasoning across product taxonomies
- **Production Architecture:** Engineered offline-online serving system implementing daily template updates and recommendation refresh cycles

Distributed Training Optimization with Ray

- **Systems Research:** Architected distributed training pipeline for a two-tower recommendation model across multi-node GPU clusters, achieving 2.7x speedup (4hr → 1.5hr on 4 GPUs)
- **Data Pipeline Innovation:** Implemented hierarchical data sharding strategy distributing GCS file reads across nodes and workers, optimizing both inter-node and intra-node data parallelism
- **Performance Analysis:** Conducted systematic bottleneck analysis and optimization, demonstrating super-linear scaling efficiency in distributed deep learning workloads

Automated Semantic Classification & Taxonomy Mapping

- **Research Problem:** Designed automated product-to-taxonomy mapping system replacing manual category assignment with individual product-level classification using Google Product Category (GPC) taxonomy
- **Embedding Architecture:** Developed feature-based product embedding system computing semantic similarity across 5,000+ GPC categories, achieving fine-grained automated classification at scale

- **Semantic Understanding:** Created representation learning framework that captures product-category compatibility relationships, enabling automated taxonomy assignment without manual rule engineering
- **Business Impact:** Deployed system improving return on ad spend (ROAS) and achieving a 10% year-over-year increase in average order value through enhanced product discoverability and targeting precision

MLOps Infrastructure & Research Support

- **Platform Modernization:** Led migration of 15+ ML pipelines from on-premises to Vertex AI, establishing standardized distributed training patterns for data science team
- **Research Tooling:** Developed comprehensive training and deployment framework using Kubeflow on GKE, enabling rapid experimentation and model iteration for data science researchers
- **Innovation Projects:** Conducted proof-of-concepts in parallel computing optimization (Ray), vector similarity search (ChromaDB), and real-time monitoring (Prometheus/Grafana)

DATA SCIENCE RESEARCH PROJECTS

Columbia University, Model Quantization using TensorflowLite (*Dec 2020*):

- Conducted systematic research on neural network compression techniques including post-training quantization, quantization-aware training, and weight pruning
- Achieved 4x model size reduction with minimal performance degradation, contributing to efficient model deployment research

Data Science Internships

Nokia Bell Labs (*Jun-Aug 2020*):

- Developed novel CNN-LSTM autoencoder architecture for multivariate time series anomaly detection, recipient of Bell Labs Innovation Award

Ralph Lauren Capstone (*Sep-Dec 2020*):

- Built return propensity prediction system using AWS SageMaker and advanced feature engineering

Ellicium Solutions (*Jan-May 2018*):

- Researched imbalanced learning techniques for customer retention in insurance domain

EDUCATION

Columbia University New York City, NY
Master of Science, Data Science Dec 2020
Relevant Courses: Machine Learning, Exploratory Data Analysis and Visualization, Probability Theory and Statistics, Statistical Inference, Algorithms.

Vishwakarma Institute of Technology Pune, India
Bachelor of Technology May 2018
Bachelor of Technology in Production Engineering, GPA: 9.44/10.

RECOGNITIONS

Publications: “Cost-Optimal Maintenance Strategies Using Machine Learning” - ORSI Conference

Awards: Nokia Bell Labs Summer Intern Innovation Award for Outstanding Research Contribution

Technical Leadership: Mentored 5+ data scientists in distributed ML deployment and MLOps best practices