Suchismit Mahapatra

MACHINE LEARNING/LLM ENGINEER · (MACHINE LEARNING | DEEP LEARNING | NLP)

Meta, Menlo Park, CA

☐ 716-435-8865 | ✉ suchismi@buffalo.edu | 🏠 schrilax.github.io/ | 🔗 suchismit | ▶ schrilaxs | ⑧ suchismit

About Me

Proficient technical leader with extensive research and developer experience in building custom, robust systems for large-scale learning and driving industry-wide impact. Proven ability of delivering complex technical initiatives from conception to production.

My research interests/expertise lies in:

- Machine/Deep Learning (ML/DL)
- Natural Language Processing (NLP/LLM)
- Deep Graph/Geometric Learning (GNN)
- Reinforcement Learning

Skills & Proficiencies

Python | PyTorch | Hive | Scala | C/C++ | TensorFlow | Keras | Apache MapReduce | CUDA

Experience

Meta *Menlo Park, CA*

MACHINE LEARNING/LLM ENGINEER *January 2024 - Present*

- Improving YA presence on comment surfaces
  - Built a comprehensive strategy for improving young adults (YA) presence on comment surfaces, resulting in tactical initiatives:
  + Developed an interesting comments based notification (serving 1.6M users daily) incorporating user personalization in VM, which resulted in strong funnel gains (clicks +0.16%, reactions +0.17%, actions +0.15%).
  + Built foundational comment-based semantic labels (e.g., p(funny), p(mean), etc.) using Llama-as-oracle which had applications in multitude of products.

- Llama-as-oracle and vibe improvement on comment surfaces
  - Tech lead (4 engineers) for LLM-in-comments work-stream.
  - Built large scale LLM inference pipelines to generate semantic labels for comments using prompt tuned *llama-3.1-70b-instruct*.
  - Utilizing above teacher model, trained highly accurate (both AUC-ROC, PR-AUC >= 0.93), "cheap" XLM(R)-based student models for serving billions of comments in production (needed to "optimally" balance compute, latency and performance).
  - Fine-tuned *llama-3.1-8b-instruct* model via SFT + LORA using a high quality dataset (1M samples) generated utilizing prompt-tuned *llama-3.1-70b-instruct* resulting in 75% GPU compute savings and performance gains (+9% PR-AUC, +15% accuracy).
  - Added "guardrail" components and incorporated best practices as part of label generation pipeline to ensure high label quality.
  - Leveraged *llama-4-maverick-17b-128e-instruct* as teacher model in large scale LLM inference pipelines to generate semantic labels for non-english comments.
  - Collaborated with central translation team for fine-tuning *llama-3.1-8b-instruct* model to further improve labeling performance.
  - Using above semantic signals in our comment ranking VM boosting funny/interesting comments + demoting bad comments etc. resulted in strong vibe gains (overall vibe +19.2%, vibe "mimicry effect" +1.19%, comment VPV +0.67%, severe bad vibe -27.3%).

LinkedIn *Mountain View, CA*

SENIOR AI SCIENTIST/ENGINEER *July 2021 - September 2023*

- Conditional label generation using LLMs
  - Built prompt generation pipelines for large scale LLM inference to assist in conditional label generation via in-context learning.
  - Worked towards instruction fine-tuning and pre-training of in-house LLMs for various tasks.

- Special Interest Group (SIG)
  - Built a novel unsupervised GNN framework which learns holistic member embeddings via incorporating edge based features in the graph convolution, which when used as seed both accelerated model training speed and improved model performance for clients.
  - Developed a novel strategy for using offline RL methods to build Task-oriented dialogue agents. 📄

- Standardization/Oribi/Groups
  - Tech Lead for Standardization team (10+ engineers), wherein worked with product managers to convert business/product requirements into practical/scalable technical solutions, applying different ML/DL, GNN and NLP techniques to solve related problems.
  - Led firefighting efforts to quickly resolve P0 issues affecting 725K+ and 183K members which resulted in $5M+ revenue gain.
  - Improved average coverage of education taxonomy from 74% to 77.2%, which measures to be +5%.
  - Built relevance-based models which significantly improved group post contributions (+19.23%) and consumption (+22.18%).

Amobee *Redwood City, CA*

SCIENTIST I *March 2020 - July 2021*

- Novel bidding strategy based on Win Price (WP) estimation
  - Developed and productionized a novel bidding strategy using nonlinear ML based approaches for estimating WP.

- Factorization Machine (FM/FFM) based ML pipeline for usage in production
  - Led efforts to build a FM/FFM based ML pipeline using a novel sparse matrix formulation that can handle high modality features.

- BERT/GAN based user embeddings
  - Incorporated BERT/GAN based user embeddings into existing ML/DL models leading to enhanced model accuracy (+7%).

Criteo AI Lab   *Palo Alto R&D Center, CA*

RESEARCH SCIENTIST   *July 2018 - December 2019*

- Improve Click-through and Sales prediction
  - Enhanced existing production Click-through and Sales prediction pipeline using nonlinear ML techniques. Improved stability of our new models significantly from +50% to +5%. A/B test using new models resulted in +3-6% uplift in long-term RexT on all platforms.
- Theoretical aspects of Deep Learning (worked with Noureddine El Karoui)
  - Working towards understanding kernel and manifold specific aspects of theoretical deep learning.
- Resolving the posterior-collapse issue in Seq2Seq learning
  - Developed a quantization based approach towards resolving the posterior-collapse issue. 📄

Criteo AI Lab   *Palo Alto R&D Center, CA*

RESEARCH SCIENTIST INTERN   *May 2017 - December 2017*

- Cross-domain Query-Product (QP) modeling
  - Developed a robust QP model across retailer domains via Domain Adaptation and Optimal Transport based approaches. 🅿

BD Biosciences   *San Jose, CA*

MACHINE LEARNING ALGORITHM DESIGN INTERN   *June 2016 - August 2016*

- Fast Clustering of Flow Cytometry (FC) data
  - Upscaled BD's clustering framework for high dimensional FC data upto ~16x. 📄🅿

Academic Background

🏛 **University of Buffalo, The State University of New York**   *Buffalo, NY*

PH.D. IN COMPUTER SCIENCE   *April 2012 - June 2018*

- Topic: Scalable Nonlinear Spectral Dimensionality Reduction methods for streaming data. 📄🅿🅿
- Advisors: Varun Chandola, Nils Napp & Jaroslaw Zola | GPA: 4.0 out of 4.0 (Transcript)

🏛 **University of Buffalo, The State University of New York**   *Buffalo, NY*

M.S. IN COMPUTER SCIENCE   *September 2010 - June 2012*

- Topic: A Cold Start Recommendation System Using Item Correlation and User Similarity. 📄
- Advisor: Rohini Srihari | GPA: 4.0 out of 4.0 | Department rank: 1 out of 555 (Transcript)

🏛 **National Institute of Technology, Rourkela**   *Rourkela, India*

B.TECH. IN COMPUTER SCIENCE   *August 2001 - May 2005*

- Specialization: Discrete Mathematics and Algorithms
- Cumulative Score: 77% (First class with Honors)(Transcript) | Joint Entrance Exam Rank 22 out of 400,000

Honors

2022	Completed NLP / NLU and RL courses as part of AI certification from Stanford University.	*Sunnyvale, CA*
2022	Was invited to and attended the prestigious 2022 CIFAR DLRL School and OxML 2022.	*Sunnyvale, CA*
2021	PC member for ICLR (2021 - present), ACL (2021 - present) and NeurIPS (2021 - present).	*Sunnyvale, CA*
2020	Was invited to and attended the prestigious Theory of Reinforcement Learning program.	*Berkeley, CA*
2019	PC member for ICML (2020 - present) and EMNLP 2021.	*Palo Alto, CA*
2019	Was invited to and attended the prestigious Foundations of Deep Learning program.	*Berkeley, CA*

Publications

1. New Methods & Metrics for LFQA tasks. **S. Mahapatra**, V. Blagojevic and P. Bertorello. 2021 (Preprint available) 📄

2. Interpretable Graph Similarity Computation via Differentiable Optimal Alignment of Node Embeddings. K. Doan, S. Manchanda, **S. Mahapatra** and C. Reddy. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021 📄

3. Discretized Bottleneck in VAE: Posterior-Collapse-Free Sequence-to-Sequence Learning. Y. Zhao, P. Yu, **S. Mahapatra**, Q. Su and C. Chen. 2020 (Preprint available) 📄

4. Learning Manifolds from Non-stationary Streaming Data. **S. Mahapatra** and V. Chandola. 2019 (Preprint available) 📄

5. S-Isomap++: Multi Manifold Learning from Streaming Data. **S. Mahapatra** and V. Chandola. Proceedings of 5th IEEE International Conference on Big Data, 2017 📄

6. Error Metrics for Learning Reliable Manifolds from Streaming Data. **S. Mahapatra**, F. Schoeneman, V. Chandola, J. Zola, N. Napp. Proceedings of SIAM Data Mining Conference, 2017 📄

7. Modeling Graphs Using a Mixture of Kronecker Models. **S. Mahapatra** and V. Chandola. Proceedings of the 3rd IEEE International Conference on Big Data, 2015. 📄

Certifications

2023	Building Generative AI Applications 📄	*FourthBrain*
2023	Prompt Engineering for LLMs 📄	*Sphere*
2022	Designing state-of-the-art Recommender Systems 📄	*Sphere*
2022	Natural Language Processing with Transformers 📄	*Hugging Face*
2022	Mastering Model Deployment and Inference 📄	*Sphere*
2022	Driving Business Impact with Machine Learning 📄	*Sphere*
2021	Building Transformer-Based Natural Language Processing Applications 📄	*NVIDIA DLI*