Soumyajit De

Last Updated on 28th March 2025

career.soumyajit.de@gmail.com | 0091-70-222-51513 | Github: github.com/lambday | LinkedIn: linkedin.com/in/desoumyajit

Education

IIT BOMBAY
MTech in Computer Science (ML)
2014 | Mumbai, IN

WB UTECH
BTech in Computer Science
2011 | Kalyani, IN

Technical Experience

DOMAINS
Click-prediction, Language Models

LANGUAGES
C++, Python, Java, SQL

LIBRARIES & TOOLS
PyTorch, ONNX, Huggingface, Keras, Pandas, SkLearn, Matplotlib, SciPy, NumPy, Jupyter, Docker, Kubernetes, Azure Data Factory, Azure Data Lake Storage, Azure Data Lake Analytics, Distributed FS (Cosmos), Map-Reduce, Kafka, BLAS, GDB, Valgrind, Perf, Git, RESTful APIs, OAuth, Conda, Pip, LaTeX.

Publications

[1] Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alexander J. Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. Link.

Honours

10th-Board Exam: State Rank: 1st
Recipient, Chief Ministers Gold Medal.

Industry Experience

MICROSOFT | Senior Data & Applied Scientist, Search Advertising R&R
Dec 2018 - Present | Bangalore, IN

ONLINE RANKING
- Introduced online ranking of all ad assets to major markets using low-latency click-prediction models utilising statistical signals. Obtained 1.5-2% ΔCTR.
- Introduced unified feature stores, trained unified models to serve across multiple clusters, extending ranking support to 100+ markets across the globe.
- Improved model by integrating textual signals from query, headline and ad asset using a multilingual encoder. Improved overall ΔAUC by 4%, helping with the cold-start problem. Addressed signal sparsity in low-volume markets through knowledge distillation and met latency demands for inference with cached embeddings. Obtained +0.3-2.0%ΔCTR across markets.

PERSONALISATION
- Performed scope analysis for coverage, proposed improvement through integrating user-profile signals in ranking, relevance, and selection for ad assets. Designed roadmap and drove the initiative from proposal to delivery working with multiple cross-functional and cross-geographic teams.
- Sourced user signals from different services across products into homogeneous text features using in-context learning. Examined user-interest clusters and explored approaches to capture diversity and recency of interests.
- Customised and fine-tuned an encoder to output lower dimensional embeddings, meeting capacity budgets while maintaining quality (0.065%→0.058% ΔAUC). Owned global integration of this approach (+0.2%ΔCTR on personalisable slice). Working on integration in online selection and relevance in an advisory role.

OFFLINE SELECTION
- Proposed, implemented and delivered an approach for offline selection of daily generated ad-assets utilising rankscore over historical queries for the ad. Worked with partner team on training a global variant of this approach. Exploited sampling to address the scale of ranking ∼10B items every day (+0.1-0.3%ΔCTR).

ASSET GENERATION
- Increased candidate density (1.5-2.0x) globally with zero-shot asset generation using an instruction-tuned encoder-decoder model (+0.22%ΔRevenue).
- Working with integration of expanded query, summarised landing page, and user profile for adaptive online asset generation in zero-shot setting. Exploring S/M/LLMs for knowledge distillation and preference optimisation utilising quality check models.

ORACLE | Senior Software Engineer, Cloud Infrastructure
Jul 2014 – Apr 2016, Sep 2016 – Dec 2018 | Bangalore, IN

- Designed and implemented a majority of the Marketplace REST API.
- Employed batch-processing and application-layer caching to reduce the response times of multi-page GET-calls from ∼2 mins to ∼10 secs.

Research Experience

UNIVERSITY COLLEGE LONDON | Research Assistant, Gatsby Unit
May 2016 – Jul 2016 | London, UK

- Devised a cache-friendly algorithm for class of statistical tests involving MMD estimator that showed ∼300x speed-up over naïve implementation.
- Proposed and implemented a multi-threaded variant that outperformed competing algorithms, built with state-of-the-art solvers, by an order of magnitude [1].

Open Source Experience

SHOGUN ML LIBRARY | Core Contributor | 94,221 LOC changes
2013, 2014, 2016 | Google Summer of Code

2016 Co-mentored in designing Shogun's Linear Algebra library.

2014 Designed and developed a framework for kernel-based hypothesis tests. Added a family of feature selection algorithms on this framework.

2013 Implemented an estimator for log-det of large, sparse matrices arising in the log-likelihood computation of high-dimensional Gaussians in real-world datasets.