Smit Hinsu

Software Engineer | ML Systems Efficiency

smittvhinsu@gmail.com
Mountain View, CA
linkedin.com/in/smit-hinsu-b6569030
+1 650-660-0121
smit-hinsu

EXPERIENCE

ML Efficiency, YouTube

Google

2023 – Present | Mountain View, CA

- Bootstrapped and led an ML efficiency workstream across 20+ teams and **100+ developers and researchers**, optimizing recommendation, generative, and safety models.
- Delivered **$40M+** in resource savings and **$100M+** in additional revenue by spearheading a multi-faceted efficiency initiative that included low-level TPU optimizations, compiler improvements, strategic workload placement, and revamping resource planning.
- Defined ML efficiency principles and fostered an efficiency-driven culture through new metrics, technical workshops, incentive alignment, and collaboration among SRE, infrastructure, and quality teams.

TensorFlow Compiler and Performance, Google Brain

Google

2018 – 2022 | Mountain View, CA

- Architected and built an MLIR-based compiler bridge for TensorFlow XLA/TPU backend & TFMobile; mentored 10+ engineers across multiple teams through design, implementation, and a seamless production rollout.
- Improved TensorFlow GPU performance by integrating TensorRT, optimizing convolution kernels, and enhancing multi-GPU performance on Google Cloud.
- Recognized as one of the top 25 contributors to the open source TensorFlow repository, driving key features and performance optimizations for the global ML community.

Indexing and Storage Infrastructure, Search

Google

2015 – 2018 | Mountain View, CA

- Architected a core lookup service for Google's indexing storage system **(5M+ QPS)**, replacing static sharding with a dynamic load balancing model that improved reliability and maintainability for 10+ critical client teams.
- Owned and delivered the first major overhaul of the indexing data format since 2008, improving data locality for 200+ internal teams; drove 1.4x to 15x performance improvements for critical data access pipelines.

Software Engineer Intern, Growth

Facebook

2013 | Menlo Park, CA

ABOUT ME

"Impact-driven Software Engineer with a decade of experience at Google building and optimizing large-scale ML systems. Fueled by a deep curiosity for the entire technology stack, from low-level hardware to compilers and modeling. Passionate about scaling impact not just through code, but by leading cross-functional initiatives, mentoring colleagues, and fostering collaborative, efficiency-focused engineering cultures."

TECHNICAL SKILLS

ML Compilers | ML Frameworks | ML Accelerators Programming | ML Infrastructure | ML Performance Optimizations | Recommendation Systems | Large-Scale Distributed Systems | Database Systems

EDUCATION

B.Tech in Computer Science

IIIT Hyderabad

2010 – 2014 | Hyderabad, India

- Led the Competitive Programming Club, organizing educational sessions, practice contests, and mentoring peers.

HONORS & AWARDS

Bonuses @ Google
Recognized with 90+ Peer and Spot Bonuses for collaborations and cross-functional leadership.

Performance Awards @ Google
Received 11 Perfy Awards for novel performance improvements.

ACM-ICPC World Finals
Ranked 1st in two Indian regionals, a first in the competition's history.