Vinu Rajashekhar

Google DeepMind

vinutheraj@gmail.com
Mountain View, California.

ML/AI Leader
Accomplished professional with over a decade of experience driving innovation and delivering results in ML/AI teams. Proven track record of leading technical teams to design, build, and scale end-to-end solutions across the entire stack, including UX, backend systems, modeling, TPUs, and on-device applications.

Seeking to lead end-to-end efforts in GenAI teams, driving the entire product lifecycle from conception to deployment. Ideally focused on features that span from fine-tuning LLMs to crafting exceptional UX.

Gemini App Team

End-to-End ML Product Leadership

- **Transformed response generation experience**: Led a 10-person cross-functional team (UX, Backend, Modeling) to overhaul response generation in the Gemini App, transitioning from single to streaming responses. Delivered **XX% increase in queries** and **X% retention growth among new users**.
- **Strategic leadership**: Convinced Gemini App VP to launch a strategic initiative to reduce Gemini App's first token latency to under 1 second, significantly enhancing user experience.
- **Innovated user engagement tools**: Led an 8-person end-to-end team to launch file upload capabilities in Gems (similar to custom GPTs), enabling personalized user interactions.

Servo Team

Building Scalable Machine Learning Services

- **Launched TensorFlow Serving**: Co-founded the Servo team to create Google's de-facto service, Servomatic, for hosting TensorFlow models, scaling from zero to **40M queries/second** (by the time I left). Open-sourced the project (github.com/tensorflow/serving), driving wider adoption and contributions outside of Google.
- **Pioneered distributed serving**: Designed a system enabling distributed TensorFlow serving, supporting massive-scale models like YouTube recommendations requiring **1.2TB memory**.

TPU Inference

Optimized Real-Time Inference for TPUs

- **Pioneered real-time TPU inference**: Led the adaption of the TPUs for real-time inference within Google, enabling low latency tasks (as opposed to high throughput tasks). Developed latency-optimized systems critical for fast product interactions on top of TPU inference, enabling **launches like Gmail SmartCompose and AlphaZero**.

Assistant NLP Team

Infrastructure & Protocol Development

- **Enhanced real-time speech understanding**: Designed a core streaming protocol that improved Google Assistant's ability to process and comprehend speech and natural language in real time.
- **Seamless integration**: Led cross-team collaborations to integrate the protocol across frameworks, ensuring robust functionality.

Sibyl Team

Scaling ML Systems for Recommendations & Predictions

- **Drove revenue and engagement growth**: Scaled Sibyl to handle 3x more launches for critical applications like ad and video recommendations, resulting in **$XB annual ad revenue** and **XX+M daily YouTube watch hours**.
- **Optimized memory efficiency**: Developed a factorized model-serving system that improved memory usage by 4x, enabling seamless experimentation and adoption across YouTube and Gmail.

Training System Optimization

- **Maximized compute efficiency**: Introduced optimizations like software prefetching, better hashmaps, more cache-friendly datastructures. Improved Sibyl's training pipeline, reducing CPU usage by **23%** and memory consumption by **33%**, saving **253,000 compute cores annually**. Significantly shortening training times and accelerating product development cycles.