Vinay Rao

Palo Alto, CA

+1-213-400-0458 • sr.vinay@gmail.com • vinaysrao.com

I am a machine learning researcher working on large language models, specifically about parameter and computational efficiency alongside scaling laws. I have previously worked on over-parameterized networks for their theoretical properties, speech recognition, and computer vision.
Programming Languages: *Proficient in*: Python,C++ *Extensively used*: C, Java, Matlab/Octave
Frameworks: PyTorch, JAX, Tensorflow

Professional Experience

Character.ai Palo Alto, CA Research Engineer August 2022-Present

- I am currently working on scaling up large language models through studying scaling laws, inference efficiency, and parameter efficient optimization.
- I have implemented deep learning systems and frameworks on several hardware accelerator families.
- At this company, I have quickly prototyped features for the website, developed PoCs on social platforms for distribution, and worked on cluster management.

Cerebras Systems Sunnyvale, CA Research Scientist Jun 2021-July 2022

- I worked on understanding the loss landscapes and optimization properties of sparse neural networks and models with limited capacity to improve parameter efficiency of networks.
- Protoyped several algorithms for inducing high sparsity in large language models.

Google Brain Mountain View, CA (Senior Research Engineer) Nov 2017-May 2021

- Researched the mechanisms of deep networks through over-parameterized networks, and extended this to normalization and finding new architectures that are theoretically motivated.
- Developed methods for using large Transformer networks for text summarization, information retrieval, and sequence-to-sequence generation.
- Publications
  - WebRED: Effective Pretraining And Finetuning For Relation Extraction On The Web
  - Is Batch Norm unique? An empirical investigation and prescription to emulate the best properties of common normalizers without batch dependence
  - A Mean Field Theory of Batch Normalization
  - Assessing The Factual Accuracy of Generated Text

Baidu Research Sunnyvale, CA (Research Scientist) Jan 2016-Nov 2017

- Developed novel architectures and algorithms for automatic speech recognition and language modeling.
- Designed scaleable architectures for character and word level language models for use in speech and dialog systems.
- Collaborated with product teams to develop a patented latency-controlled recurrent architectures for deployable speech recognition models.
- Developed normalization and optimization techniques for recurrent networks.
- *Publications*
  - Deep speech 2: End-to-end speech recognition in english and mandarin
  - Active Learning for Speech Recognition: the Power of Gradients
  - Reducing Bias in Production Speech Models

Aindra Systems, Bangalore, India (Research Engineer) Jan 2013-Jul 2013

- Developed algorithms for an automated attendance system with face recognition and tracking.
- Implemented the entire product stack including the website, APIs, and mobile app.
- Prototyed a system for automatic detection of cancerous cells through imaging.

Amazon, Bangalore, India (Software Engineer) Aug 2013-Jun 2014

- Developed a large scale real-time product and vendor reporting tool.
- Built an easily configurable floating ad banner system for mobile websites.
- Worked on a secure sign-in page for mobile and creating data-stores and aggregators for search queries.

Academic Experience

Robotics and Embedded Systems Laboratory, C.S Dept, University of Southern California May 2015 - Dec 2015.

Graduate Student Assistant

Systems, algorithm development, simulations and backend work for autonomous aerial and aquatic vehicles.

- Built a multi-view adaptable object tracking system for aquatic vehicles.
- Developed an in-flight camera simulator for aerial autonomous vehicles.

Data Analytics Laboratory, E.E Dept, University of Southern California May 2015 - August 2015

Graduate Student Research Assistant

Computer vision, statistics and deep learning for medical imaging data (MRI, fMRI).

- Developed a novel deep learning architecture for segmenting tumorous cells in MRI images for BRATS (Brain tumor segmentation challenge) 2015.
- Researched several ways to perform multi-modal learning and stacking to achieve high recall rates for tumor types.
- *Publications*
  - Brain tumor segmentation with deep learning, MICCAI 2015

Master's thesis USC, CA, USA 2015

On the optimization techniques in high-dimensional clustering, dimensionality reduction and visualization

- Extensively surveyed state of the art algorithms for unsupervised learning such as Stochastic Neighbor Embedding, Spectral Clustering, Word2Vec and Auto-encoders and compared their results in the domains of clustering and visualization.
- This study included comparison of run-times, optimization techniques and implementation of the algorithms in the study.

Bachelor's Thesis BMSCE, Bangalore, India 2013

A holistic view on object recognition in videos

- Comprehensive survey and study of historic to state of the art algorithms and features for generic object recognition in videos.
- Implemented several algorithms including multinomial regression, Linear SVMs, and some feature extractors.
- Presented comparative results of recognition with hand-crafted (SURF/SIFT) features against CNNs for real-time recognitin and localization in videos.

Education

- **University of Southern California** — **Los Angeles**
  M S Computer Science, GPA:3.51/4.0 — *Aug 2014-December 2015*
  Courses: Advanced Algorithms, Artificial Intelligence, Convex and Combinatorial Optimization
  Probabilistic Reasoning, Brain Theory and Artificial Intelligence, Computer Vision
- **Visvesvarayya Technological University** — **Bangalore**
  B S Computer Science, GPA: 8.78/10.0 — *Sep 2009- May 2013*
  Courses: Pattern Recognition, Probability & Statistics, Advanced data structures and algorithms, Networks, OS, Compilers