Bhargav Chippada

Website: https://bhargavchippada.github.io/
408-674-9633 | bhargavchips@google.com | bhargav.chippada19@gmail.com
852 E Dana St, Mountain View, CA, 94041
https://www.linkedin.com/in/bhargavchippada/

Education

University of Massachusetts Amherst — Amherst, MA
Master of Science in Computer Science with concentration in AI — ***Aug 2019 - Dec 2020***
Associated with the BINDS lab (Biologically Inspired Neural & Dynamical Systems) under Prof. Robert Kozma

Indian Institute of Technology (IIT), Bombay — Mumbai, IN
Bachelor of Technology in Computer Science and Engineering, ***All India Rank 47*** — ***Jul 2012 - May 2016***

Relevant Courses

UMass Courses:
- Advanced Machine Learning
- Reinforcement Learning
- Systems for Data Science
- Quantum Information Systems
- Robotics
- Probabilistic Graphical Models
- Math Tools for Data Science
- Secure Distributed Systems
- Advanced Algorithms

- Convolutional Neural Networks for Visual Recognition
- Statistics 110: Probability. Machine Learning. Artificial Intelligence

Open Source

Published and Maintaining Popular Python Package for ForceAtlas2 algorithm on PYPI *2017*
- The fastest Python implementation on PyPi under the alias "fa2". Received more than **300 stars** on GitHub
- Used by researchers to visualize huge graphs in fields like AI, Physics, and Molecular Biology
- The algorithm simulates a physical system with the nodes of the graph as charged particles and edges as springs
- Implemented Barnes-Hut approximation to improve the time complexity from $O(n^2)$ to $O(n.\log n)$
- GitHub: https://github.com/bhargavchippada/forceatlas2

Published Faster Whisper for Offline Dictation on PYPI (Privacy Focused) *2025*
- Real-time speech-to-text dictation powered by faster-whisper. Speak and watch text appear instantly in any application – fully offline, no cloud APIs, no data leaves your machine
- GitHub: https://github.com/bhargavchippada/faster-whisper-dictation

Project Cuecard - The right rule, at the right moment *2026*
- Contextual rule enforcement for AI coding agents. Cuecard retrieves your most relevant guidelines and injects them before every tool call and user message — so the agent always follows your rules
- Multi-stage RAG pipeline: multi-retriever (dense + sparse) → LLM (opt-in) → Rules for hooks
- GitHub: https://github.com/bhargavchippada/cuecard

Humanizing AI workshop, IJCAI conference *Stockholm, SE, Jul 2018*
- Presenter and first author of a poster titled *"Knowledge Amalgam: Generating Jokes and Quotes Together" - arXiv*

Experience

Google — Mountain View, California
Software Engineer, Senior Machine Learning — *January 2021 - Present*
Google Lens & Multimodal Search Experience — *C++, Python, Tensorflow, Gemini*
- Lead the design and deployment of Multimodal DeepRank for Lens image-only queries from training to serving
- DeepRank is a deep learning model that's a core component of Google's search ranking system for Web Results
- Teacher-student distillation to a 6-layer funnel transformer to predict click and information satisfaction for relevance
- Increased CTR on Web Resutls by a significant +2% globally and Information Satisfaction of 9% Lens queries
- Leading technical strategy, end-to-end development, optimization, and launch of core deep learning models handling millions of multimodal query requests per minute while working cross-functionally with multiple stakeholders
- Mission: Building Multi-Modal search experience in Google to enable users to search using both image and text

Video Understanding for Shopping Videos (Previous Team)
- Identified prominent products reviewed in a shopping video using its metadata like title, description, ASR, and OCR
- Built ML pipelines that run daily on fresh videos to extract the products and show it on Google Search page

Google Mountain View, California

ML Software Engineer Intern *May 2020 - August 2020*

AV1 Codec (Predictive Video Compression) — *C++, Python, Tensorflow, Convolutional Neural Networks*

- Used CNN to restore the quality of degraded frames after its encoded using inter-frame coding technique
- Used spatial separable convolutions to reduce the number of neural network parameters by 85%
- Achieved −1.2% reduction in BD-rate for the same PSNR compared to the state of the art with just 10k parameters
- Advanced video compression technology to achieve better video quality on lower network bandwidths

Microsoft R&D Pvt Ltd Bangalore, IN

Software Engineer *Mar 2018 - Jul 2019*

Azure Partner Analytics Team — *Spark, Scala, SQL, Python, Azure*

- Migrated the Xbox360 purchases' data processing pipeline to Azure cloud services without any downtime and issues
- Integrated real-time purchases data into reports thereby cutting down the delay from 12 hours to 15 mins

Bing Local Search Team — *CosmosDB, Scope/SQL, C#, Python, Keras, PyTorch*

- Improved the business listings ranking by using an effective popularity measure for the business entities
- Used signals like search counts, impressions, click counts, ratings, and reviews to predict a target popularity score

Ola Cabs (ANI Technologies) Bangalore, IN

Software Engineer *July 2016 - Mar 2018*

Location Intelligence Team — *Java, C++, Python, Docker, AWS, Spark, PostGIS, OSM*

- Implemented HMM based map matching algorithm for snapping the noisy GPS updates of a cab to the correct road
- Built using Open Street Routing Machine in C++ to ensure speed while handling millions of requests per minute
- Improved the efficiency of Ola's tracking system by 20%
- Developed OlaMaps, an internal custom map tile server, and website for geo-visualizations
- Worked on time-dependent routing (source to destination navigation), route optimization, and mining geospatial data
- Implemented a scalable community detection algorithm for weighted graphs in Spark

Amazon Development Centre Bangalore, IN

Software Engineer Intern *May 2015 - July 2015*

Consumer Marketing Analytics Team — *Java, Hadoop, Apache Pig*

- Credited purchases of customers to different marketing channels using various attribution models on click-stream data
- Implemented standard models like last interaction, time decay, linear and position based in an extensible manner

Knit Messaging, College Startup Mumbai, IN

Software Engineer — Android *March 2014 - July 2014*

- Co-developed the Android App for one-way information dissemination between a teacher and parents of their students

Academic Distinctions

- Ranked 47 in IIT-Joint Entrance Examination among 500,000 students in 2012
- Ranked among top 35 students of India by HBCSE Olympiad examinations (Physics, Astronomy, and Science)
- Received Young Scientist Award (KVPY fellowship) for ranking among top 0.01% students in India, 2011
- Ranked 1 in Dr A.S Rao Science Examination of Andhra Pradesh, India, 2010
- Teaching Assistant for the Computer Programming and Utilization course, Spring semester, IIT Bombay in 2016
- Elected technical secretary of Hostel-3 during the 2013-14 academic year
- Grader for the Quantum Information Systems course, Fall semester, UMass-Amherst, 2020

Skills

Programming Languages: C++, Python, Java, Scala, Go, SQL, HTML & Javascript, LaTeX
Libraries: PyTorch, Tensorflow, Fastai, PostGIS, ReactJS
Technologies: Spark, Docker, Android, Web, Linux, AWS, Azure, Google Cloud
AI Tools: Claude, Codex, Gemini, RAG, Agentic Workflows