Sravanth Kodavanti

Bengaluru, Karnataka, 560048, India

+91-9550173710 ramasravanthkodavanti@gmail.com sravanthk27 Google Scholar OpenReview Website

About

Research Engineer specializing in **pre/post-training optimization** and **efficient AI** for real-world impact across **language** and **vision** models. Skilled in advanced **inference optimization**, **model compression** (quantization), and **Neural Architecture Search**, delivering performance improvements in both industrial and research projects. Passionate about bridging cutting-edge AI methods with scalable solutions to address practical business needs.

Education

Indian Institute of Technology, Hyderabad **2020 - 2024**
Bachelor of Technology, **Major** - Computer Science and Engineering, **Minor** - Entrepreneurship CGPA - 8.58

Research Publications

1. Subhajit Sanyal*, Srinivas Miriyala*, Akshay Bankar*, Manjunath Arveti, Sowmya Vajrala, Shreyas Pandith, ***Sravanth Kodavanti***, Abhishek Ameta, Harshit, Amit Unde
NanoSD: Edge Efficient Foundation Model for Real Time Image Restoration,
Under Review at CVPR 2026

2. ***Sravanth Kodavanti****, Srinivas Miriyala*, Sowmya Vajrala*, Vikram N R, Sharan Allur
Edge-Efficient Image Restoration: Transformer Distillation into State-Space Models,
Under Review at CVPR 2026

3. Sowmya Vajrala*, Srinivas Miriyala*, ***Sravanth Kodavanti***
Towards Efficient Image Deblurring for Edge Deployment, *Preprint*

4. ***Sravanth Kodavanti****, Sowmya Vajrala*, Srinivas Miriyala*, Utsav Tiwari* *et al*
Unlocking the Edge Deployment and On-Device Acceleration of Multi-LoRA Enabled One-for-All Foundational LLM, *Preprint*

5. Srinivas Miriyala*, Sowmya Vajrala*, Hitesh Kumar, ***Sravanth Kodavanti***, Vikram N R
Mobile-friendly Image De-noising: Hardware-Conscious Optimization for Edge Application,
ICASSP 2025

Patents

1. ***Sravanth Kodavanti****, Srinivas Miriyala*, Sowmya Vajrala*
System and Method for Accelerating On-Device Large Language Model Inference.
Provisional Specification Filed. Proposes a novel self-speculative decoding technique to significantly improve efficiency and reduce latency in LLM inference on edge devices.

Work Experience

Samsung Research Institute Bangalore **Aug 2024 - Present**
Machine Learning Research Engineer ***Mentor: Dr. Srinivas Miriyala***

- Developing **Samsung Neural Acceleration Platform** for AI model acceleration and deployment on mobile devices, leveraging **Neural Architecture Search** (NAS) , **Quantization** and other techniques for performance optimization.
- **Commercialization & Impact:**
  1. **Accelerated inference speed** or the **Samsung LLM Gauss - L, 3B/3.5B Model** by implementing
     * **Speculative Decoding**, achieving a **2X improvement** in tokens per second (toks/sec), Successfully integrated into the **Samsung S25 series**.
     * Developed **PULSE**: a novel in-house **self-speculative decoding** algorithm, achieving a **5X improvement** in tokens per second (toks/sec). Method targeted for **Samsung S26 series**.

2. Developed **NAS and quantization** driven models delivering advanced, commercialization-ready imaging solutions, including:
   * UNet-based model for **low-light video de-noising** (**2.5**× speedup; deployed on **Samsung Galaxy A56**),
   * UNet-based model for **demoire artifact removal** (**2.2**× speedup; targeted for **Galaxy S26**),
   * Stable Diffusion-based model for **text-to-image generation, motion photo enhancement, and related use-cases** (**2.3**× speedup; targeted for **Galaxy S26**)

- **Awards & Recognition:**
  1. **Spot Award (Q2 2025)** – For novel speculative decoding algorithm enhancing LLM acceleration.
  2. **MD Project Incentive Award (2024–25)** – For AI model optimizations and deployment on edge devices.
  3. **Team Awesome Award (Q4 2024)** – For on-device optimization of **Samsung Gauss - L** in **S25** mobile series.

Stealth Startup - Gika Graph.AI **June 2024 - Sep 2024**
AI Researcher ***Mentor: Dr. Manoj Aggarwal***

- Developed domain-specific search engines leveraging **Knowledge Graphs** (KG) to enhance semantic accuracy. Improved **Retrieval Augmented Generation** by using KG over vector databases for more precise, context-aware retrieval.
- Handled **Coreference** and **Entity Resolution** for fine-tuning data using LLMs such as **GPT-4, GPT-4o**, **LLAMA3**, **SpanBERT**, and **LingMess** with the **SpaCy** AI framework.

Hexagon R & D India **Jan 2024 - June 2024**
Machine Learning Intern ***Mentor: Ankan Sengupta***

- Developed a website for implementing segmentation & classification based on tags of various manufacturing plant sketches.Used various segmentation algorithms such as **RANSAC**, **DBSCAN**, **K-Means**.
- Used **Azure form recognizer** models & other OCR models for tag identification.
- Developed a **flask website** for text recognition on manufacturing plant sketches using **Form, Doc Recognizer** & also used OCR models for tag identification.
- Involved in an **LLM** research project. Compared the results for text generation between **Mixtral**, **Mistral**, **LLAMA2**.

OnePlus / Oppo (OPLUS) Mobiles India R & D **Jan 2023 - June 2023**
Research Intern - Device AI ***Mentor: Dr. C Shyam Anand***

- Implemented model compression techniques called **Quantization** , **Pruning** & **Distillation** on various deep learning models such as **ResNet** , **Yolo** , **ViTs** , **ConvNeXt** , **Stable Diffusion Models** & **Large Language Models** (LLAMA) . **Compressed all these models for deploying in mobile devices.**
- Involved in research work on **Neural Style Transfer** (NST).
- Impemented a DL model for **LaTeX - OCR** task . The model's aim is to detect & recognize the mathematical equations present in a research paper. Compressed the model for the integration with edge devices.

Teaching Experience

Department of Computer Science , IIT Hyderabad **Nov 2021 - Apr 2024**
Teaching Assistant

- Served as a **Teaching Assistant** for multiple courses, including:
  1. **Operating Systems** (CS3510) under ***Dr. Sathya Peri***
  2. **Discrete Mathematics** (CS1010) under ***Dr. Rakesh Venkat***
  3. **Database Management Systems** (CS3550) under ***Dr. Manish Singh***
  4. **Theory of Computation** (CS2030) under ***Dr. Subrahmanyam Kalyanasundaram***

Research Community Service

Research Reviewer **2025 – Present**

- **Reviewer**: **ICML**, **NeurIPS**, **ICLR**, **ACL**, **AAAI** & **ICASSP** workshops and main tracks.
- **Area Chair**: **ICASSP'26**

Projects

Continual Learning for 3D Point Cloud ***Guide: Dr. P K Srijith***

- Implemented Continual Learning on **Pointnet** architecture by addressing the challenge of catastrophic forgetting.
- Implemented **Knowledge Distillation** approach for the Continual Learning & Model was trained on **ModelNet10** dataset.

Computer Vision & NLP | Personal Projects

- Image denoising using **Auto - Encoders** . Model trained on **MNIST** dataset.
- Human Face generation using **GANs** . Model trained on **celeba** dataset.
- Underwater Object Detection and Classification using **DGYOLO**. Model trained on **URPC 2019** dataset.

Technical Skills

Languages: C, C++, Python, JavaScript

Frameworks & Tools : PyTorch, Tensorflow, ONNX, Hugging Face, Docker, Git, LaTeX, Markdown

HPC Frameworks: CUDA, Triton, DeepSpeed, TensorRT, ONNX Runtime

Achievements

- Rank **1156** in IIT JEE Advanced 2020 among **150,000** participants.
- Rank **493** in JEE Main B Planning 2020 among **59,000** participants.
- **99.75** percentile in IIT JEE Main 2020 among **1.1 Million** participants.
- AP EAMCET 2020 Rank **135** among **156,000** participants.
- TS EAMCET 2020 Rank **307** among **143,000** paricipants.
- Solved around 700+ CP problems over multiple platforms LeetCode , CodeChef , CodeForces , GeeksForGeeks.

Leadership / Extracurricular

- Worked as a Core member for **Epoch : AI-ML club of IITH**.
- Worked as an **Internship** & **Placement** Coordinator for Office of Carrer Service of IITH.
- Member of IITH **Chess** team , Represented IITH at **InterIIT** & various competitions.