Ravisri Valluri

linkedin.com/in/ravi99 | ravi-vk.github.io | US Citizen

Objective

Researcher specializing in large-scale retrieval, machine learning for low-latency search, and NLP-driven user intent understanding. Interested in efficient machine learning, natural language processing, representation learning, generalization, reasoning, and multimodal AI.

Education

Indian Institute of Technology, Madras — Bachelor of Technology **2023**

GPA: 9.51/10.00 | Major: Computer Science and Engineering

Guru Junior College, Hyderabad — Intermediate Board XII **2019**

Grade: 95.8% | Stream: Math, Physics and Chemistry

Publications

- ***Scaling the Vocabulary of Non-autoregressive Models for Fast Generative Retrieval***
  Ravisri Valluri, Akash Kumar, Kushal Dave, Amit Singh, Jian Jiao, Manik Varma, Gaurav Sinha
  Accepted at **KDD 2025, Research Track**.

Selected Projects

Foundation Retrieval Models — Dr. Manik Varma, Microsoft Research **July 2023 – Present**

- Introduced a new approach to generative document retrieval, achieving **70x faster GPU inference** using non-autoregressive (NAR) models compared to autoregressive (AR) models, while maintaining competitive or improved accuracy.
- **Scaled vocabularies to 5 million tokens**, 100x larger than BERT's 50K tokens, integrating multi-word entities and phrases to reduce token dependencies and enhance NAR model performance, achieving improved retrieval accuracy and closing the gap with AR models.
- **Optimized inference** for large vocabulary retrieval models by application of machine learning techniques like self-normalization for efficient SoftMax approximation and a novel learned shortlisting method to reduce the search space for beam search.
- **Accepted at KDD 2025** and presented at the ICML 2024 SPIGM workshop, with patents pending.

Visual Question Answering — Prof. Chandra Sekhar, IIT Madras **Jan 2023 – May 2023**

- Developed an architecture for **open-ended Visual Question Answering** (VQA), using a transformer encoder to project images into the embedding space of GPT-2, enabling the model to generate answers and explanations for multimodal queries despite being trained only for text.
- Grounded inputs with external knowledge graphs like ConceptNet, improving the model's ability to perform complex visual reasoning by **incorporating structured world knowledge**.
- Designed a three-stage pretraining process utilizing image captioning datasets to transfer learning to VQA and other vision-language tasks, compensating for more limited VQA datasets.

Experience

Microsoft Research, India — Research Fellow **July 2023 – Present**

- Developed an approach for low-latency, accurate generative retrieval models by scaling vocabularies and optimizing inference algorithms, deployed in collaboration with Microsoft Advertising across 150+ markets, generating **tens of millions** in revenue.
- Initiated and drove the **Extreme Vocabulary Project**, which grew from a small team into a significant research initiative recognized by leadership, including Mustafa Suleyman, and presented at the MSR Academic Summit to leading researchers in Indian academia.
- Wrote and implemented code for training and inference of **multi-billion parameter** models using advanced deep learning frameworks, handling datasets containing **trillions of tokens**.
- Currently working on further **scaling vocabularies** and **improving training techniques** for efficiency, accuracy, and robustness of large vocabulary language models.

Microsoft Development Center, India — Software Engineering Intern **May 2022 – July 2022**

- Developed an **ML inference system** for seamless task/model switching, building and deploying a REST API using Python, Flask, SQL, and Azure Machine Learning.

Digital Outcomes, Remote — Summer Intern **May 2021 – July 2021**

- Built a **Hyperledger Fabric network** to track the provenance of items in a supply chain, implemented smart contracts, and developed application logic using Java.

Technical Skills

- **Programming Languages & Tools:** Python, C++, SQL, Java, Flask
- **ML Frameworks & Platforms:** PyTorch, ONNX, DeepSpeed, HuggingFace, Azure ML Platform
- **ML Techniques & Algorithms:** large language models, extreme classification, inference & decoding algorithms, tokenization, sampling techniques, distillation, contrastive learning, graph machine learning, approximate nearest neighbors search, quantization

Scholastic Achievements

- Recipient of the S. Subramanian Prize in the 2021 Institute Day awards for the **highest CGPA (9.98/10.00)** out of all B. Tech and Dual Degree first year students at IIT Madras.
- Secured an All-India Rank of **897 out of 1.1M** students in JEE MAIN, 2019.
- Secured an All-India Rank of **699 out of 160K** shortlisted students in JEE Advanced, 2019.

Relevant Coursework

- **AI/ML:** Fundamentals of Deep Learning, Pattern Recognition and Machine Learning, Natural Language Processing, Advances in the Theory of Deep Learning.
- **CS:** Discrete Mathematics, Programming and Data Structures, Design and Analysis of Algorithms, Randomized Algorithms
- **Systems:** Foundations of Computer System Design, Computer Architecture and Organization, Compiler Design, Operating Systems, Introduction to Database Systems
- **Math:** Multivariable Calculus, Series and Matrices, Basic Graph Theory, Probability, Statistics and Stochastic Processes