Yerram Varun

vyerram@google.com | varun221.github.io | LinkedIn | Google Scholar | GitHub

Education

Indian Institute of Technology, Guwahati *2019 - 2023*
Bachelor of Technology in Electronics & Communication
Minor in Computer Science & Engineering GPA: **8.5**/10.0

Research Experience

Google DeepMind, India *Aug 2023 - Present*

Pre-Doctoral Researcher
Advisors: Dr. Prateek Jain , Dr. Praneeth Netrapalli & Dr. Karthikeyan Shanmugam

Time Reversed LLMs

- Implemented pre-training & post-training pipelines with token reversal data processing to train Reversed-LLMs of **LaMDA**, **PaLM2** and **Gemini** Family of models.
- Performed exhaustive experiments into various applications viz Reranking, Citations, Retrieval, Safety and Planning.
- Showed extensive improvements in Instruction following abilities, Citation attribution and Passage Retrieving capabilities of Gemini-1.0 Pro resulting in **Spotlight submission @ NeurIPS, 2024**.

Efficient Inference: High Recall Estimation

- Worked on efficient inference in FFN and Softmax layers of LLM through **sparsity exploitation** techniques.
- Performed extensive ablation studies to design algorithms that **cheaply** predict top non-zero rows/columns of activations with **high recall** in distributed inference settings.
- We achieve **1.47x** per step decode speedup with minimal finetuning on 1 billion parameter models.

Long Context Efforts : Gemini Family of Models

- Worked on approximate logit computation and clustering approaches to identify top-k keys in the attention layer.
- Designed approaches to retain quality and improve latency with existing methods in a distributed inference setting.
- Showed significant latency improvements on Gemini at different scales and settings.
- Actively contributing to the productionization efforts to improve current Long-context Gemini models.

FlowVQA: Multimodal Reasoning in LLMs

Collaborator: Prof. Vivek Gupta

- Developed a QA Benchmark on flowcharts to test reasoning capabilities of Multimodal LLMs.
- Designed and generated four different types of Question sets: (a) Topological (b) Flow referential (c) Scenario & (d) Fact Retrieval. Evaluated leading open-source and closed-source LLMs on this benchmark.
- This work was accepted in the **Findings of ACL 2024.**

Indian Institute of Technology, Guwahati *Jan 2022 - Mar 2022*

Collaborator: Prof. Vivek Gupta

Trans-KBLSTM: Enhancing Tabular Natural Language Inference

- Developed custom architectures using **RoBERTa** and **BiLSTMs** to integrate external knowledge bases like **ConceptNet** and **Wordnet** into attention architectures.
- Improved SOTA Accuracy on Infotabs, a tabular Natural Language Inference Dataset by 4-5% on low-resource settings.
- Published and received **Best Paper Award** at DeeLIO Workshop@ACL 2022

Publications

- **Time-Reversal Provides Unsupervised Feedback to LLMs.**
  Y Varun*, R Madhavan*, S Addepalli*, A Suggala, K Shanmugam, P Jain.
  Neural Information Processing Systems, 2024.
  Spotlight Submission.
- **Does Safety Training of LLMs Generalize to Semantically Related Natural Prompts?**
  S Addepalli, **Y Varun**, A Suggala, K Shanmugam, P Jain.
  SafeGenAI Workshop @ Neural Information Processing Systems, 2024
  Under submission to ICLR, 2025
- **HiRE: High Recall Approximate Top-k Estimation for Efficient LLM Inference**
  Y Samaga B L*, **Y Varun***, C You, S Bhojanapalli, S Kumar, P Jain, P Netrapalli

- **Trans-KBLSTM: An External Knowledge Enhanced T-BiLSTM Model for Tabular Reasoning**
  Y Varun, A Sharma, V Gupta
  DeeLIO Workshop @ Association for Computational Linguistics, 2022
  🏆 **Best Paper Award**.
- **FlowVQA: Mapping Multimodal Logic in Visual Question Answering with Flowcharts**
  S Singh*, P Chaurasia*, **Y Varun**, P Pandya, V Gupta, V Gupta, D Roth
  Findings of Association for Computational Linguistics 2024

Engineering Experience

India Machine learning Team, Amazon Science, Bangalore *May 2022 - July 2022*
Applied Scientist Intern

- Worked on (i) Top-K feature-selection based and (ii) Temporal learning based attention architectures to model customer demographics through Amazon Behavioral history.
- Improved the existing XGBoost production pipeline by 2-3% F1 score with **Time2Vec** and **Multi-head Attention** based algorithms and contributed to initial efforts to productionize.

RethinkUX, India *Jan 2021 - Feb 2021*
DL Software Development Intern

- Developed **text detection models** and **text recognition methods** to extract information from PDF documents.
- Implemented parsing-based pipelines to extract required customer attributes from parsed documents.
- Created a **Flask API** and used **Docker** to containerise the application for easy deployment over system.
- Created a complete **ML pipeline** from data preparation to deployment in production.

Achievements

- **Peer Bonus** , For providing guidance on coding infrastructure & Model/Data workflows @ Google DeepMind, India
- **Peer Bonus** , For providing guidance on dataset pipelines @ Google DeepMind, India
- **Best Paper Award DeeLIO Workshop, Association for Computational Linguistics 2022**

Hackathons

- **Amazon Machine learning Challenge, 2021**, National Runner up among 10,000+ registered participants
- **Inter IIT Tech Meet 10.0**, Secured Silver Medal in Bosch Model Extraction Module
- **Inter IIT Tech Meet 9.0**, Secured Silver Medal in the Bosch Traffic Sign Recognition Module

Kaggle ML Competitions

- **RSNA 2022 Cervical Spine Fracture Detection - Kaggle**, Finished 38th (Silver Region), Top 5%
- **Mayo Strip AI Competition - Kaggle**, Finished 27th (Silver Region), Top 4%
- **DFL : Bundesliga Data Shootout - Kaggle**, Finished 44th (Silver Region), Top 8%
- **Sartorius Cell Instance Segmentation - Kaggle**, Finished 111th (Bronze Region), Top 8%
- **Mechanisms of Action - Kaggle**, Finished 208th (Silver Region), Top 5%

Relevant Coursework

Mathematics - Linear Algebra, Multivariate Calculus, Probability & Random Processes(AA Grade)
Computer Science - Deep Learning (AA Grade), Data Structures & Algorithms, Machine Learning, Computer Vision

Technical Skills

Languages (Python, C/C++, MATLAB, LaTeX, HTML/CSS, JS, Bash, SQL, MATLAB), *Libraries* (JAX, Pytorch, Tensorflow, Keras, FastAI, PySpark, Flask, Django), *Tools* (Git, HuggingFace, Bootstrap)

ExtraCurriculars

- **Secretary, IITG.ai**, Responsible for managing events, workshops, competitions, and discussions at IITG.ai club, the AI Community of IIT Guwahati
- **Speaker at MLTechFest' 21** , Conducted a Workshop on basics of transformers at MLTechFest'21 organized by Tensorflow User Group (Mysuru) & Google Developer's Group (Mysuru)
- **Computer Vision Workshop by IEEE SFIT & IEEE APSIT** , Conducted an online workshop session on basics of Computer Vision and OpenCV
- **First Position in Robothon 2020 organized by Robotics Club, IITG**
- **First Position in Astronomy Quiz, Interhostal Technical Competition, Kriti 2020, IITG**