Rahul Chand

+91-650-509-9551 | rahulsc@stanford.edu | chandrahul0320@gmail.com
Website: rahulschand.github.io | **Linkedin:** /RahulSChand | **GitHub:** /RahulSChand

EDUCATION

Stanford University **2024-2026**
Masters in Computer Science

Birla Institute of Technology, Pilani **2015-2019**
Bachelor in Computer Science | 9.68 GPA

INDUSTRY AND RESEARCH EXPERIENCE

Microsoft Research | Research Fellow **July 2021-Nov 2023**

- Worked with the Turing Team on Transformer compression for Microsoft Office products.
- Developed factorization and sparsity methods to reduce memory footprint (~4x compression with <1% drop in GLUE/SQuAD + internal benchmarks). Submitted to EMNLP 2022.
- Worked with Manik Verma and Yashoteja Prabhu on Extreme Multi Label Learning (XML) to improve robustness of Bing's Ad recommender systems.
- My work primarily involved studying & improving tail performance of extreme classifiers via distillation (5% improvement over SOTA one-vs-all classifiers). Work published at ICLR'24.

Arcesium | Software Engineer & Intern **Aug 2019-May 2021**

- Worked in the Performance and Accounting team as a full stack developer.
- Worked as part of the team responsible for developing microservices & frontend using Java, Kotlin, Python, ReactJS & T-SQL for handling large volumes (>100k) of trades daily.

Indian Institute of Science (IISC) | Undergraduate Thesis student **Jan 2019-July 2019**

- Worked with Dr. Ram Prabhakar at Video Analytics Lab(VAL) for my undergrad thesis on capsule networks.
- I developed methods to study and use capsules for optical flow estimation. Also released an open-source library to generate toy datasets for optical flow. Work submitted to ICCV 2019 Neural Architects workshop.

Indian Institute of Remote Sensing (IIRS) | Research Intern **May 2017-July 2017**

- Worked with the Geo-informatics Department on their road-asset management project.
- I developed a car tracking system for Dehradun roads using kalman filter and haar-like features. Also worked on road-asset mapping of Indian roads using Faster-RCNN finetuned on images of Dehradun roads (report).

PUBLICATIONS

Peer-reviewed conference proceedings

- Enhancing Tail Performance in Extreme Classifiers by Label Variance Reduction (**ICLR 2024**) | Paper Link
  *Anirudh Buvanesh, **Rahul Chand (co-first author)**, Yashoteja Prabhu, Manish Gupta*

Workshop and arxiv papers

- DSFormer: Effective Compression of Text-Transformers by Dense-Sparse Weight Factorization (arxiv link)
  ***Rahul Chand**, Yashoteja Prabhu, Pratyush Kumar*.
- CapsFlow: Optical Flow Estimation with Capsule Networks (arxiv link)
  ***Rahul Chand**, Rajat Arora, Ram Prabhakar, Venkatesh Babu*.

OPEN SOURCE CONTRIBUTIONS

- **Compute requirements for LLMs** (Github) **| 1000*+ stars***
  - Tool to check GPU vRAM requirement and token/s for training & inference of any LLM. Supports frameworks like HuggingFace, vLLM, llama.cpp, and quantization (bitsandbytes, GGML & QLoRA).
  - The tool has been used over **130k+ times** by **20k+** users.
- **llama2.c for dummies** (Github) **| *200+ stars***
  - Walkthrough tutorial of llama2.c written as a starter reference for LLM inference.

- **llama2.c** (Github) **|** ***12k+ stars*** **(contributor)**
    - A vanilla, barebones implementation of the LLaMA language model in C, trained on the TinyStories dataset; developed by Kapathy.
- **Fast & tiny datasets for optical flow (**Github**)**
    - Library to generate tiny optical flow datasets on the fly for sanity testing optical flow estimation models. Written as part of undergraduate thesis at IISC & used in the paper link.
- **Efficient Batched Torch KSVD** (Github)
    - Library to run sparse dictionary completion algorithm KSVD on batched matrices on GPU. Written using pytorch as part of transformer compression work at Microsoft Research.
- **Attention network for reading comprehension and question answering** (Github)
    - Tensorflow implementation of the paper "Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering."

TEACHING EXPERIENCE

Teaching Assistant for below six courses. Graded assignments, prepared course projects & supervised lab sessions.

- **Fall 2018:** Data Mining, Principles of Programming Languages, Computer Programming
- **Spring 2018:** Data Structures and Algorithms, Database Systems
- **Fall 2017:** Logic in Computer Science

ACHIEVEMENTS

- One of 30 students selected from Maharashtra (population of 100+ million) to attend training camp for INMO (Indian National Maths Olympiad) 2015.
- Received the Merit-cum-Need scholarship at BITS Pilani for all eight semesters, awarded to students with excellent academic record and financial need.

SKILLS

Programming Language: Python, C, C++, Java, JavaScript, Kotlin

Libraries and Frameworks: PyTorch, Numpy, TensorFlow, Keras, HuggingFace, ReactJS, NextJS