Saurabh Raje

(+91) 8003140887 | ✉ saurabh.mraje@gmail.com | ⌂ smr97.github.io | ⌨ smr97 | ⦿ saurabhmraje

Education

Birla Institute of Technology and Science, Pilani *Pilani, India*
BE (Hons.) in Computer Science and Engineering *Aug. 2015 - Dec. 2018*

Delhi Public School, Gurgaon *Gurgaon, India*
Senior secondary school *2013 - 2015*

Work Experience

IBM Research *Delhi, India*
Research Engineer
Scaling AI *August 2019 - Present*

- Working with the Model compression team to make AI faster.
- Worked with Watson NLU product team to reduce inference time of BERT for entities tagging.
- Contributed to the implementation of *PowerBERT*, a new model that is **4.5x faster** than **BERT** for inference.
- This work was published in **ICML**'20, and was integrated into IBM OneNLP product stack.
- Worked with biotechnology researchers to build a fast bulk RNA sequencing pipeline for **COVID-19** research.
- Currently working on large scale training of Temporal Graph Convolutional Networks (**GCN**) to detect money laundering.
- Co-invented 2 patents on model compression techniques and multiobjective optimisation.

ETH Zurich *Zurich, Switzerland*
Scientific Assistant
High Performance Computing for Deep Learning *March 2019 - August 2019*

- Worked on accelerating training process of Deep Neural Networks using the **DACE** language developed in-house.
- DACE is a domain specific language for HPC workloads that uses a novel Stateful Dataflow Graph (SDFG) based Intermediate Representation.
- Many graph based transformations can be done on this SDFG IR, and generate fast code for heterogenous platforms.
- Wrote a Tensorflow frontend for DACE that parses a TF computation graph to build a DACE SDFG.
- Achieved at-par performance for ResNet-50 in comparison to TF and CUDNN.

INRIA *Grenoble, France*
Bachelor Thesis
Parallel Programming *September 2018 - February 2019*

- Developed **Kvik**: a task based middleware in the **Rust** language.
- **Kvik** provides tunable task splitting strategies that can be composed with each other.
- Wrote a state-of-the art parallel sort implementation using Kvik, that scaled up to 64 threads.
- It offered 1.65x more speedup than GNU parallel sort, and 2x more speedup compared to C++20 ParallelSTL sort with Intel TBB.
- This work has been submitted to **PPoPP**'21.

IBM Research *New Delhi, India*
Research Intern
Scaling AI *May 2018 - July 2018*

- Implemented the **GoogleNet** with variable batch sizing and activation checkpointing.
- This led to a 20% reduction in training time under memory constraints.
- Also worked on various optimisation heuristics for decomposition of sparse tensors.
- These outperformed best known heuristics in literature by 30%.

BITS Pilani *Pilani, India*
Research Assistant
Parallelizing Compilers *August 2017 - December 2018*

- Worked on the DWARF domain specific language compiler developed in-house.
- The compiler parallelizes code for various data mining applications, written in the DWARF language.
- Contributed to the modelling of various data dependencies in the context of density based and hierarchical clustering algorithms.
- Finally deployed a new optimisation layer that increased the granularity of parallelism.
- The system hence achived linear speedup for DBSCAN, SNN, and RECOME clustering algorithms.

UST global

Trivandrum, India

Research Intern

June 2017 - August 2017

Deep Learning for cybersecurity

- Developed a malware detection engine using a deep belief network (DBN).
- Achieved an accuracy of 89.1 and true positive rate of 98.2.
- This project was presented to Cyberdome, Kerala Police and led to a MoU between UST Global and Cyberdome for future research.

Team Anant - A nanosatellite development team

Pilani, India

Team lead, on board computing

January 2016 - January 2018

Embedded systems

- Contributed to the development of an on board computer for a nanosatellite.
- Lead a group of ten students to this effect.
- Built a fault tolerant software to run complex monitoring and control algorithms for the satellite.
- Several device drivers for the Linux kernel were built from scratch to interface sensors and actuators on the satellite bus.
- The satellite will be launched by the Indian Space Research Organisation.

Honors & Awards

2020	**Winner**, Awesome Team Award - IBM India Research Lab
2018	**Winner**, Best Poster Award - IBM India Research Lab
2017	**Winner**, Mercedes Benz Hack.Banglore 2018
2016	**Winner**, Best Paper Award - APOGEE (BITS Pilani's technical festival)

Presentations

Mobile World Congress 2018

Barcelona, Spain

Presenter for Daimler AG

February 2018

- Invited by Daimler AG to present our winning hackathon prototype at the MWC.
- The prototype was built to detect pedestrians using low cost IR sensors.
- This would allow for level 4+ automated driving.

Skills

Languages	Rust, Python, C, C++, Java
Frameworks	PyTorch, Tensorflow, Caffe, CuDNN, Git
HPC Libraries	openMPI, openMP, Intel TBB

Publications

[1] V. J. Badami, K. Aggarwal, S. Sharma, S. M. Raie, and T. Goyal. In-loop simulation of attitude control of a nanosatellite. In *2019 IEEE Aerospace Conference*, pages 1–9. IEEE, 2019.

[2] V. T. Chakaravarthy, S. S. Pandian, S. Raje, and Y. Sabharwal. On optimizing distributed non-negative tucker decomposition. In *Proceedings of the ACM International Conference on Supercomputing (ICS)*, pages 238–249, 2019.

[3] S. Goyal, A. R. Choudhury, S. M. Raje, V. T. Chakaravarthy, Y. Sabharwal, and A. Verma. Power-bert: Accelerating bert inference via progressive word-vector elimination, 2020.

[4] S. Islam, S. Balasubramaniam, P. Goyal, A. Sultana, L. Bhutani, S. Raje, and N. Goyal. A rapid prototyping approach for high performance density-based clustering. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 260–269. IEEE, 2019.

[5] S. Raje, A. Goel, S. Sharma, K. Aggarwal, D. Mantri, and T. Kumar. Development of on board computer for a nanosatellite. *68th International Astronautical Congress (IAC)*, 2017.

[6] S. Raje, S. Vaderia, N. Wilson, and R. Panigrahi. Decentralised firewall for malware detection. In *2017 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pages 1–5. IEEE, 2017.