Saurabh Raje

(+1) 801-706-7032 | saurabh.raje@utah.edu | smr97.github.io | smr97 | Saurabh Raje | saurabhmraje

Education

University of Utah
Salt Lake City, USA

PhD in Computer Science and Engineering

Aug. 2021 - Present

- Advised by Prof. Saday Sadayappan

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 - August 2021

- Worked with the Model compression team to make AI faster.
- Implemented *PowerBERT*, a new model that is up to **4.5x faster** than **BERT** for inference.
- This work was published in **ICML'20**, and was integrated into IBM OneNLP product stack.
- Co-developed a new method to train Dynamic Graph Neural Networks on supercomputers.
- Collaborated with biotechnology researchers to build a fast bulk RNA sequencing pipeline for **COVID-19** research.
- Co-invented 4 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.
- Wrote a Tensorflow frontend for DACE that parses a TF computation graph to build a DACE SDFG.
- Added a pattern based kernel fusion transformation on the DACE SDFG to reduce kernel calls and repetitive memory access.
- 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.
- Revamped a visualisation tool to measure the *speed* of each task, the number of tasks created and stolen.

IBM Research
New Delhi, India

Research Intern

Scaling AI

May 2018 - July 2018

- Worked on training deep networks under memory constraints.
- Implemented variable batch sizing coupled with activation checkpointing for the *GoogleNet*.
- This led to a 20% reduction in training time under memory constraints.
- Implemented various optimisation heuristics for the decomposition of sparse tensors.
- FLOPs reduction was achieved by reordering the Tensor-Times-Matrix product in the Tucker decomposition algorithm.
- Memory usage was reduced with a Compressed Sparse Fibre representation for sparse Tensors with mixed-mode ordering.
- These outperformed best known heuristics in literature by up to 30%. Research was published in **ICS'19**.

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 generates parallel code with MPI calls for various data mining applications.
- Modelled the data dependencies for density based and hierarchical clustering algorithms.
- Built a new optimisation layer that increased the granularity of parallelism.
- The generated code 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.
- Delivered this work to the Inspector General of Cybercrime at Kerala Police.

Team Anant

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

2021 **Winner,** Patent Plateau Award - IBM India Research Lab

2021 **Winner,** Outstanding Technical Achievement Award - IBM India Research Lab

2020 **Winner,** Distinguished Paper Award - IBM India Research Lab

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, **Raje, Saurabh**, 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, **Raje, Saurabh**, 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, **Raje, Saurabh**, V. Chakaravarthy, Y. Sabharwal, and A. Verma. PoWER-BERT: Accelerating BERT inference via progressive word-vector elimination. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 3690–3699, Virtual, 13–18 Jul 2020. PMLR.

[4] S. Islam, S. Balasubramaniam, P. Goyal, A. Sultana, L. Bhutani, **Raje, Saurabh**, 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] A. Kannan, A. Roy Choudhury, V. Saxena, **Raje, Saurabh**, P. Ram, A. Verma, and Y. Sabharwal. Hyperaspo: Fusion of model and hyper parameter optimization for multi-objective machine learning. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 790–800, 2021.

[6] **Raje, Saurabh**, 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.

[7] **Raje, Saurabh**, 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.

[8] **Saurabh Raje** and F. Wagner. Kvik: A task based middleware with composable scheduling policies, 2020.

[9] Y. Xu, **Raje, Saurabh**, A. Rountev, G. Sabin, A. Sukumaran-Rajam, and P. Sadayappan. Training of deep learning pipelines on memory-constrained gpus via segmented fused-tiled execution. In *Proceedings of the 31st ACM SIGPLAN International Conference on Compiler Construction*, CC 2022, page 104–116, New York, NY, USA, 2022. Association for Computing Machinery.