Amey **Agrawal**

CS Ph.D. Student, Georgia Tech

ameya.info @ agrawalamey12@gmail.com github.com/agrawalamey Google Scholar

Education

Present
Aug 2022
Georgia Institute of Technology **Atlanta, USA**
Ph.D., Computer Science. GPA 4.00/4.00.
Advisor: Prof. Alexey Tumanov | Area: Systems for machine learning, LLM inference systems.

Jul 2018
Aug 2014
Birla Institute of Technology and Science Pilani **Pilani, India**
B.E. (Hons.), Computer Science

Experience

Present
May 2023
Microsoft Research **Remote**
Research Intern | Mentor: Dr. Ramchandran Ramjee
Designing efficient inference systems for large language models.

Aug 2022
Jan 2021
Microsoft Research **Bangalore, India**
Research Software Engineer-II | Mentors: Dr. Muthian Sivathanu
Built parts of the elasticity sub-system that leveraged efficient time sharing of GPUs to provide transparent scaling of deep learning training workloads. This work was done as a part of the Singularity project, Microsoft's planet-scale AI infrastructure service.

Nov 2020
Jul 2018
Qubole Inc. **Bangalore, India**
Member of Technical Staff-II | Mentors: Rohit Karlupia
Worked on various applied machine learning and software engineering problems to enhance Qubole's data science platform. Published research in several top-tier venues.

Dec 2017
Jul 2017
Software Engineering Intern | Mentor: Bharath Bhushan
Built core data-plane components for Qubole's Deep Learning clusters based on TensorFlow and Apache Spark.

Publications

Sarathi: Efficient LLM Inference by Piggybacking Decodes with Chunked Prefills [pdf]
Amey Agrawal, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, Ramachandran Ramjee
Preprint: arXiv:2308.16369 (2023) **[CoRR]**

Vidur: A Large-Scale Simulation Framework For LLM Inference
Amey Agrawal, Nitin Kedia, Jayashree Mohan, Ashish Panwar, Nipun Kwatra,
Bhargav S. Gulavani, Ramachandran Ramjee, Alexey Tumanov [Coming Soon]

DynaQuant: Compressing Deep Learning Training Checkpoints via Dynamic Quantization [pdf]
Amey Agrawal, Sameer Reddy, Satwik Bhattamishra, Sarath Nookala,
Vidushi Vashishth, Kexin Rong, and Alexey Tumanov
Preprint: arXiv:2306.11800 (2023) **[CoRR]**

Sybill: Deep Learning Workload Tuning with Virtual GPUs [poster]
Srihas Yarlagadda, **Amey Agrawal**, Sarath Nookala, Pranavi Bajjuri, Shivam Mittal, Alexey Tumanov
ACM Symposium on Cloud Computing Poster, 2023 **[SoCC'23]**

Singularity: Planet-Scale, Preemptible, Elastic Scheduling of AI Workloads [pdf]
Singularity Team, Microsoft
Preprint: arXiv:2202.07848 (2022) **[CoRR]**

Logan: A Distributed Online Log Parser [pdf]
Amey Agrawal, Rajat Gupta, and Rohit Karlupia
Proceedings of IEEE International Conference on Data Engineering, 2019, Macau **[ICDE'19]**

Learning Digital Circuits: A Journey Through Weight Invariant Self-Pruning Neural Networks [pdf][code]
Amey Agrawal, and Rohit Karlupia
Sparsity in Neural Networks Workshop 2021; New in ML Workshop, NeurIPS, 2019, Vancouver [**SNN'21**]

Delog: A Privacy Preserving Log Filtering Framework for Online Compute Platforms [pdf] [dataset]
Amey Agrawal, Abhishek Dixit, Namrata Shettar, Darshil Kapadia,
Rohit Karlupia, Vikram Agrawal, and Rajat Gupta
Proceedings of IEEE International Conference on Big Data, 2019, Los Angeles [**BigData'19**]

Honours and Awards

Center for Research into Novel Compute Hierarchies (CRNCH) Fellowship, 2023 [🌐]

- For research of automatic hardware-aware optimization of deep learning training workloads.

Teaching

Systems for Machine Learning *Teaching Assistant | Prof. Alexey Tumanov* Fall'23

- Conducted a three-part lecture series on large language model inference systems.

Introduction to Neural Networks & Fuzzy Logic *Lead Teaching Assistant | Prof. Surekha Bhanot* Fall'17, Spring'18

- Introduced Python programming assignments along with a new custom-built evaluation platform. Other responsibilities included coordinating the team of seven teaching assistants to conduct labs, designing assignments, and helping students with the term project.

Introduction to Machine Learning *Teaching Assistant | Prof. Kamlesh Tiwari* Spring'18

- Conducted introductory sessions on the scientific Python ecosystem, and organized tests and programming assignments for over 100 students in the class.

References

Prof. Alexey Tumanov	*Assistant Professor, Georgia Tech* [🌐]
Prof. Kexin Rong	*Assistant Professor, Georgia Tech* [🌐]
Dr. Ramchandran Ramjee	*Partner Research Manager, Microsoft Research* [🌐]
Dr. Muthian Sivathanu	*Distinguished Scientist, Microsoft Research* [🌐]
Dr. Bhargav Gulavani	*Principle Research Engineer, Microsoft Research* [🌐]