Aditya Desai

Aditya.P.Desai@rice.edu linkedin webpage +1-346-391-7488

Education

2019 – 2024 **Ph.D., Rice University, Houston, TX**
Research Interests : I am currently interested in the following areas: **Foundations of next generation AI models** – how to make AI exponentially cheaper? The approach, I am taking, is to organically combine learnings from probabilistic methods to exploit strong presence of power law in ML to achieve huge improvements. I am expanding my research to adjacent areas such as **FlexDeploy**: Solving the extreme deployment challenges for large scale machine learning including LLMs and VLMs.
CGPA : 3.92/4 (overall) 4/4 (computer science courses), A+ in research related courses

2009 – 2013 **B.Tech. Computer Science and Engineering, IIT Kanpur, India**
Thesis title *Program synthesis using natural language*. Developed natural language based instruction system for various domains such as air travel query, document editing, etc.
CGPA : 9.6/10

Research Publications

Published

1. A. Desai, K. Saedi, W. Apoorv, J. Lee, K. Zhou, and S. Anshumali, "Accelerating Inference with Fast and Expressive Sketch Structured Transform," in *(To appear in) Advances in Neural Information Processing Systems*, 2024.
2. A. Desai and A. Shrivastava, "In defense of parameter sharing for model compression," in *The Twelfth International Conference on Learning Representations*, 2024.
3. A. Desai, K. Zhou, and A. Shrivastava, "Hardware-Aware Compression with Random Operation Access Specific Tile (ROAST) Hashing," in *International Conference on Machine Learning*, PMLR, 2023, pp. 7732–7749.
4. Z. Liu, A. Desai, F. Liao, *et al.*, "Scissorhands: Exploiting the Persistence of Importance Hypothesis for LLM KV Cache Compression at Test Time," in *Advances in Neural Information Processing Systems*, 2023.
5. A. Desai, L. Chou, and A. Shrivastava, "Random Offset Block Embedding (ROBE) for compressed embedding tables in deep learning recommendation systems," in *Proceedings of Machine Learning and Systems*, D. Marculescu, Y. Chi, and C. Wu, Eds., **Outstanding Paper Award**, vol. 4, 2022, pp. 762–778.
6. A. Desai and A. Shrivastava, "The trade-offs of model size in large recommendation models: 100GB to 10MB Criteo-tb DLRM model," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., 2022, pp. 33 961–33 972. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/dbae91512889255613 4f1c5375855590-Paper-Conference.pdf.
7. Z. Dai, A. Desai, R. Heckel, and A. Shrivastava, "Active sampling count sketch (ASCS) for online sparse estimation of a trillion scale covariance matrix," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 352–364.
8. A. Desai, Z. Xu, M. Gupta, A. Chandran, A. Vial-Aussavy, and A. Shrivastava, "Raw Nav-merge Seismic Data to Subsurface Properties with MLP based Multi-Modal Information Unscrambler," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 8740–8752. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/498f2c21688f6451d9f5fd09d53edda7-Paper.pdf.

9. A. Desai, S. Gulwani, V. Hingorani, *et al.*, "Program synthesis using natural language," in *Proceedings of the 38th International Conference on Software Engineering*, 2016, pp. 345–356.
10. A. Desai, E. Jain, and S. Roy, "Facilitating Verification in Program Loops by Identification of Static Iteration Patterns," in *2013 20th Asia-Pacific Software Engineering Conference (APSEC)*, IEEE, vol. 1, 2013, pp. 83–90.

Selected Preprints

1. A. Desai, B. Meisburger, Z. Liu, and A. Shrivastava, "Heterogeneous federated collaborative filtering using fair: Federated averaging in random subspaces," *arXiv preprint arXiv:2311.01722*, 2023.
2. A. Desai, Y. Pan, K. Sun, L. Chou, and A. Shrivastava, "Semantically Constrained Memory Allocation (SCMA) for embedding in efficient recommendation systems," *arXiv preprint arXiv:2103.06124*, 2021.

Under Submission

1. A. Desai, S. Sonkar, S. Anshumali, and R. Baraniuk, "DNA : Diagonal Normalized Attention Alleviating self-bias in the self-attention mechanism of Transformer models."
2. A. Desai, T. Zhang, G. Gupta, and S. Anshumali, "IDentity with Locality: An ideal hash for efficient gene sequence search."
3. T. Zhang, A. Desai, G. Gupta, and S. Anshumali, "HashOrder: Accelerating Graph Processing Through Hashing-based Reordering."

Employment History

Aug 2024 - Current — **Postdoctoral Researcher, University of California, Berkeley** Working on futuristic LLM systems with focus on efficiency.

2022 - 2023 — **PinLabs Researcher,** Pinterest Inc, Palo Alto, Caifornia
Improving the model compression of embedding tables in recommendation models using similarity based memory sharing.

2022 — **Machine Learning Intern,** Pinterest Inc, Palo Alto, Caifornia
Improving recommendation models at Pinterest using large embedding tables served via compression - improved engagements metrics by 10% in offline evaluation.

2013 – 2019 — **Strategist,** Tower Research Capital LLC, Gurugram, Haryana, India
I worked as a part of Asian Equities market (India, Taiwan, Korea and Japan) trading team responsible for an array of different high frequency trading strategies in stocks, stock futures and index futures.
Responsibilities:
- Manage a multi-million-dollar portfolio of stocks and stock futures in Indian Markets
- Set up and maintain Post Trade Analysis system for Asian Equities
- Research new signals and devise new strategies in Asian markets.

Awards and Achievements

2023-24 — Future Faculty Fellow, 2023-2024

Awards and Achievements (continued)

- 2022-23 Ken Kennedy Fellowship, 2022-2023
- 2022 MLSYS Outstanding Paper
- MLSYS Travel Grant, 2022
- 2021 Top 10% reviewers at ICML, 2022
- 2019-20 Awarded Pollard Fellowship at Rice University for the year 2019-2020
- 2012 ACM ICPC Kanpur site Regional Finalist
- Globally 1st position at an international event 'Chaos', Techkriti
- Globally 26th position and 7th in India amongst 800+ international teams at IOPC, Techkriti
- 1st position at Kodefest, Takneek
- 2011-12 Awarded "The Certificate of Merit for Academic Excellence" in IIT Kanpur
- 2011 Awarded CBSE Mert Scholarship Scheme.
- Awarded "Best Poster Presentation" at SURGE
- Awarded "Students-Undergraduate Research Graduate Excellence (SURGE)" grant
- ACM ICPC Kanpur Site Regional Finalist
- 2010-11 Awarded "The Certificate of Merit for Academic Excellence" in IIT Kanpur
- 2010 Awarded CBSE Mert Scholarship Scheme.
- 2009-10 Awarded "The Certificate of Merit for Academic Excellence" in IIT Kanpur
- 2009 All India Rank 96 in IIT Joint Entrance Examination 2009
- All India Rank 269 and state rank of 30 in AIEEE, 2009
- Awarded CBSE Mert Scholarship Scheme.

Academic Activities

Teaching and Mentoring

- Fall 2023 **ML Efficiency Reading Group** Organized research reading group for undergraduate, masters and graduate students. [ ]
- Summer 2023 **ML Efficiency Reading Group** Organized research reading group for undergraduate and masters students to introduce them to research. [ ]
- **Research Experience for Undergraduate Mentor**, Mentored an undergraduate for Summer Google REU program. The mentee received **Best Poster Award.**
- Spring 2023 **Instructor**, COMP 600 presentation coach.
- Spring 2022 **Coursework Assistant**, Developed problem sets for "Discrete Maths" for online masters program at Rice University
- Summer 2021 **Research Experience for Undergraduate Mentor**, Mentored two undergraduates for Summer Google REU program.
- Spring 2021 **Teaching Assistant**, Probabilistic Algorithms and Datastructures
- Fall 2020 **Teaching Assistant**, Algorithms and Datastructures

Industry Talks

- 2023 *New methods for model compression*, **NVIDIA**
- 2022 *Compressed embedding tables for recommendation models*, **Google**
- *Compressed embedding tables for recommendation models*, **Meta**
- *Compressing recommendation models*, **Intel and Criteo**

Academic Activities (continued)

- *Compressing recommendation models*, **Pinterest Inc.**

2021
- *Beyond Convolutions: A Novel Deep Learning Approach for Raw Seismic Data Ingestion*, **Shell Corporation**

Workshops and Conferences

2024
- Reviewer for Neurips, KDD, ICLR

2023
- Co-organizer of Research on Algorithms & Data Structures (ROADS) to Mega-AI Models Workshop, MLSys 2023
- Reviewer for Neurips, ICLR

2022
- Reviewer for ICML, Neurips, ICLR

Technical Skills

Coding	C, C++ (openmp, mpi, cuda), Python, Java, C#
Learning and data science	pytorch, tensorflow, pandas, scipy, matlab