Kabir Nagrecha

knagrech@ucsd.edu | +1 (626) 348-6704

Education

2021-2024 PhD in Deep Learning Systems, University of California San Diego

- Meta PhD Fellowship
- Jacobs School of Engineering Fellowship
- Halicioglu Data Science Institute Fellowship

2019-2021 B.Sc. in Computer Science, University of California San Diego

- GPA: 4.0

Professional Experience

2023 Machine Learning Platform Research Intern (Netflix)

- Designed GPU kernels for accelerated self-attention operations in Transformers. Achieved 4.5X speedups over existing implementation without quality loss.
- Built first internal Stable Diffusion training infrastructure over hundreds of GPUs.
- Build first LLM serving infrastructure at Netflix — to be used for high-throughput use cases, including conversational recommendations.

2022 Machine Learning Platform Research Intern (Netflix)

- Designed a system for scaling out data parallel recommender model training at Netflix. Improved multi-GPU scaling by a factor of 16-20X on 8 GPUs versus prior infrastructure.
- Improved core recommendation metrics by 1.5% *without an architecture redesign.*
- System is used to build >100 new recommender model instances *daily*, reflecting a 16X improvement in experimentational velocity. Projected annual savings of over $8M.

2021 Machine Learning Intern (Apple)

- Redesigned Siri model development pipeline with Apache Spark and PyTorch to enable faster and more accurate training.
- Used ASR pruning techniques to reduce end-to-end running time by more than 85% while increasing accuracy by up to 40%. Results will be used to build production models and enable locale expansions.
- Projected annual savings of over $10M.

2020 Machine Learning Intern (Apple)

- Primary contributor on the Apple Watch "Raise-to-Speak" model overhaul. Developed prototype system using TensorFlow and Keras that reduced false-reject-rate of Siri invocation on Apple Watches by more than 64%, while reducing model complexity and power draw by 50%.
- System was deployed to production, and is used more than 40 million times a week by Apple Watch users around the world.
- One of 5 finalists in Apple's internal product proposal and development contest.

Awards and Fellowships

2022 META Ph.D. Research Fellowship

2022 ACM Undergraduate Student Research Competition Grand Finalist

2021 ACM SIGMOD Undergraduate Student Research Competition Winner

2021 UCSD CSE Excellence in Research Award

2021 UCSD Jacobs School of Engineering Fellowship

2021 UCSD Halicoglu Data Science Institute Fellowship

2021 CRA Outstanding Undergraduate Researcher Honorable Mention

2020 UCSD Outstanding Undergraduate Researcher Honorable Mention

Publications and Articles

2024 **InTune: Deep Reinforcement Learning for Allocating Resources over Recommender Pipelines**
Kabir Nagrecha, Lingyi Liu, Pablo Delgado
ACM TORS 2024 **(Invited Paper)**

- Paper invited to Transactions on Recommender Systems as part of the "Highlights of RecSys" issue.
- Expansion of previous work on RL for Dataloaders to cover multi-node allocations, GPU allocations, & more.

2023 **Routing Over LLMs Using Proxy Metrics for Relative Quality Estimation**
Kabir Nagrecha, Arun Kumar, Hao Zhang
Under submission at MLSys 2024

- Proposes a novel approach to estimating relative LLM quality for a given query without needing to approximate challenging evaluation metrics.
- Automatically routes queries over candidate LLMs, boosting overall serving throughput by 2-3X with only a marginal quality loss.

2023 **Saturn: Resource-Aware Multi-Query Optimization for Multi-Large-Model Deep Learning Workloads**
Kabir Nagrecha, Arun Kumar
VLDB 2024

- Auto-selects hybrid parallel training strategies, allocates GPUs over jobs, and builds an optimized multi-model schedule.
- First cluster management tool to unify all critical aspects of large-model training.
- Collaborators at Netflix are now using it to train LLMs for content production.

2022 **InTune: Deep Reinforcement Learning for Automatic Dataloader Pipeline Resource Allocation**
Kabir Nagrecha, Lingyi Liu, Pablo Delgado, Prasanna Padmanabhan
RecSys 2023

- A deep reinforcement learning-based system to automatically allocate CPU resources across dataloader pipeline stages for ML workloads.
- Finds near-optimal resource allocations quickly for any ML workload, including DLRM and Transformer training.
- Built and tested in collaboration with Netflix. Now being used for critical production data-loading pipelines.

2022 **Hydra: A System for Large Multi-Model Deep Learning**
Kabir Nagrecha, Arun Kumar
(Preprint)

- A new platform for model parallelism leveraging multi-model execution and mixed memory hierarchies to improve scalability and parallelism for training massive deep learning models.
- Enables training of multi-billion parameter Transformer models on a single 8GB GPU.
- Speeds up training by >7.5X with 8 GPUs, and demonstrates up to 50% speedups versus pipeline parallelism.
- Collaborators are now using to train 3D convolutional networks for high-resolution fluid simulation prediction.

2021 **Systems for Deep Learning**
Kabir Nagrecha
The Gradient

- Describes the evolution of the new ”MLOps” space, and how systems research is affecting the development of machine learning pipelines.

2021 **Model Parallel Model Selection for Deep Learning Systems**
Kabir Nagrecha
ACM SIGMOD 2021

- **ACM UG Student Research Competition Winner**
- Competition abstract proposing a platform for model parallelism leveraging novel ideas from RDBMSs to redesign machine learning training.

2021 **Gradient-based Algorithms for Machine Teaching**
Pei Wang, Kabir Nagrecha, Nuno Vasconcelos
CVPR 2021

- A technique for dataset pruning and selection allows a machine learning algorithm to determine most useful samples of a dataset for use in teacher-student training or human education.
- Demonstrates up to 47% increase in human student accuracy and a 140% improvement on simulated workers.

2020 **Incremental and approximate computations for accelerating deep CNN inference**
Supun Nakandala, Kabir Nagrecha, Arun Kumar, Yannis Papakonstantinou
ACM TODS 2020 **(Invited Paper)**

- System for incremental computation of repeated convolutional inference requests, reusing information between queries.
- Developed the video-inferencing adaptation to enable more efficient deep learning inference on edge-devices.

2020 **Cerebro: A Layered Data Platform for Scalable Deep Learning**
Arun Kumar, Supun Nakandala, Yuhao Zhang, Side Li, Advitya Gemawat, Kabir Nagrecha
CIDR 2021 (Vision Paper)

- A platform for model selection using new forms of mixed task-data parallel training that enables practitioners to evaluate hyperparameters and architecture for their task quickly.
- Currently being used by medical researchers.

Talks and Presentations

2023 Distributed Infrastructure for LLMs & Conversational Systems (Netflix Research Seminar)

2023 Stable Diffusion at Scale (Netflix Research & Studio ML)

2023 Hydra for LLMs (FAIR)

2023 InTune: Deep Reinforcement Learning for Automatic Dataloader Pipeline Resource Allocation (RecSys 2023)

2023 Saturn: Democratizing Fine-Tuning of Open-Source Large Language Models via Joint Systems Optimization (ODSC West 2023)

2023 Systems for Deep Learning at Scale (Clear Ventures)

2023 Systems for Deep Learning at Scale (Thesis Proposal Talk)

2022 Automatically Scaling out Data Parallel DLRM Training (ML Platform Meetup)

2022 Scaling out DLRM Training Pipelines across CPU, GPU, and Memory (Netflix Machine Learning Platform Research Talk)

2021 Optimizing ASR Lattice-RNN Training Pipelines (Apple Machine Hearing Research Seminar)

2021 Applications of Systems Research in Machine Learning Pipelines (Apple Siri Research Seminar)

2021 Hydra: Model Parallel Model Selection for Deep Learning Systems (ACM SIGMOD SRC - 1st place)

2021 Machine Teaching for Dataset Production (CVPR)

2020 Deep Learning for Multi-Model Data in Mobile Environments (Apple Siri Research Seminar)

2020 Model Parallelism and Data Management (UC San Diego Databases Lab)

2020 Machine Learning for Tunnel Excavation (University of Colorado, Boulder)

2020 Using Deep Neural Networks to Predict Sensor Response and Geology Ahead of a TBM (Intelligent Systems for Transportation Tunnel Analysis Webinar)

Technical Skills

- Programming Languages: *Python, CUDA, Scala, SQL, C, C++, Java, JavaScript*
- Data Platforms: *Spark, Hadoop, Hive, PostgreSQL*
- Machine Learning Frameworks: *PyTorch, TensorFlow, Keras, Triton, SciKit-Learn*
- Cloud Tools: *AWS EC2, Google Cloud Platform, Kubernetes, Docker, OpenAI APIs*
- Data Analytics Tools: *Pandas, NumPy, SciPy, OpenCV*