Anant Shyam

914-406-6790 | ais64@cornell.edu | https://anantshyam.github.io/ |

Education

Cornell University Ithaca, NY

Bachelor of Science in Computer Science (Honors), Minor in Mathematics *Aug '21 – May '25*

- GPA: 3.742
- Graduate Level Coursework: Advanced Machine Learning Systems, Foundations of Reinforcement Learning, Matrix Computations, Applied Functional Analysis (planned), Convex Analysis (planned), Parallel Computing (planned)

Research Experience

Cornell University Relax ML Lab Ithaca, NY

Machine Learning Researcher, Advisor: Prof. Chris De Sa *Dec '24 – Present*

- Project: Serving Pre-trained Large Language Models with Compute-Efficient Layers
- Investigating the effectiveness of distilling large LLMs into smaller models with structured layers. Expressing the linear layers of the original model in terms of structured Tensor Train (TT) and Block Tensor Train (BTT) matrices to reduce the number of trainable parameters. When doing distillation, we optimize the reverse KL divergence between the logits of the bigger and smaller models. Using GPT2-120M with dense layers as the larger model, and our GPT2-120M with structured TT/BTT layers as the smaller model.

Cornell University Computer Systems Lab Ithaca, NY

Machine Learning Researcher, Advisor: Prof. Zhiru Zhang *Jan '24 – Present*

- Project: Accelerating Large Language Model Inference on Associative Processing Units (APU)
- Optimizing low precision (4-bit) matrix multiplication on APU in order to allow for these computations to be parallelized across multiple cores to reduce time for data transfer between the APU and the host CPU. Implemented an efficient ReLU activation layer on the APU, whose total time for data movement and computation was just about 88 ms, for high dimensional (4096 x 4096) dimensional inputs.
- Developing mathematical models to predict the total latency of a data movement methods as a function of the parameters, without actually compiling and running the program. From these models, we realized that the data movement operations are a large bottleneck. Some potential enhancements to get even better performance include high bandwith memory (HBM). Planning to develop quantization algorithms to reduce the overall memory consumption of LLM, without compromising on model accuracy.

Teaching Experience

Cornell University Department of Computer Science Ithaca, NY

Teaching Assistant/Consultant *Aug '22 – Present*

- Introduction to Reinforcement Learning (CS 4789/5789) - Spring 2025
- Systems Programming (CS 4414) - Fall 2024
- Introduction to Analysis of Algorithms (CS 4820) - Summer 2024
- Data Structures and Functional Programming (CS 3110) - Fall 2023, Spring 2024
    - **Nominated for the Cornell Computer Science Course Staff Exceptional Service Award for my work as a teaching assistant in Spring 2024. I was among the top 10% of the course staff.**
- Object-Oriented Programming and Data Structures (CS 2110) - Spring 2023
- Introduction to Computing using Python (CS 1110) - Fall 2022

Industry Experience

MathWorks Natick, MA

Engineering Development Intern *May '24 – Aug '24*

- Developed an AI-based pricing assistant using MATLAB which generates the correct code to price the financial documents of potential customers. Supported European and Bermudan options, Callable Bonds, and Fixed Bonds. Presented my pricing assistant to the MathWorks Financial Modeling Team, and explained its' potential for integration with the MathWorks MATLAB Financial Instruments Toolbox.
- Conducted research to gauge the effectiveness of large language models to improve ease-of-use for classic computational finance workflows such as translating econometric models and formulating portfolio optimization problems.

Cleo Rockford, IL

Software Engineer Intern *Jun '23 – Aug '23*

- Developed a machine learning model using Tensorflow which learns how to write code in a hypothetical object-oriented language similar to Cleo's language with accuracy of approximately 60%. Pitched my model to the Director of Product Development and explained its' potential to be expanded to learn how to program in Cleo's programming language.
- Trained multiple existing models to write code that handles complex data transformations, similar to those that Cleo supports. Pitched these trained models to the Product Development Team, and explained their potential to reduce the room for error and time needed for Cleo to handle data transformations.
- Implemented a feature on the Cleo Integration Cloud, the UI where all of Cleo's customers' jobs are displayed, which allows for direct navigation from a customer's jobs to another customer's jobs using TypeScript, Angular, and RxJS.

Interactive Brokers Greenwich, CT

Software Engineer Intern *Jun '22 – Aug '22*

- Enhanced the tool that the Risk Team used to approve or reject margin changes for equities to reduce room for error and time required to review these margin changes.
- Retrieved margin-rates related data (specifically the margin rule input) from numerous relational databases using MySQL. Stored the margin rule input using Perl, and properly displayed the data for the customer on the UI using HTML, allowing the customer to make more informed decisions about the stocks that they want to purchase.
- Developed a Python script which interprets margin-rates related data including the type of stock, asset classes, and the registration date from a CSV file, and updates the appropriate relational database with this data. Incorporated an argument parser, allowing the user to pass in additional stock-related parameters to be put into the database from the command line.

Academic Reports

- **Accelerating Newton's Method using Krylov Subspace Methods** (Authors: Anant Shyam, Boao James Chen). Assessed various Krylov Subspace methods to accelerate the convergence of Newton's method on convex settings (used the a9a dataset with binary logistic regression loss). Analyzed the effectiveness of second order methods in non-convex settings (trained a deep neural network on the CIFAR10 dataset using a low rank Hessian matrix formed with largest eigenvalues and corresponding eigenvectors).
- **Accelerating the Convergence of Policy Gradient Methods** (Authors: Anant Shyam, Daniel Cao). Analyzed various existing policy gradient (PG) methods, such as momentum-based PG methods and PG methods with variance reduction, in terms of their convergence to optimal policies in a reinforcement learning (RL) setting.

Awards and Recognitions

- Cornell Outstanding Computer Science Course Staff Award - Nominated for this award for my work as a Teaching Assistant for Data Structures and Functional Programming (CS 3110). Was among the top 10% of course staff for CS 3110.
- Cornell Engineering Dean's List Scholar
- Honorable Mention for Best Overall App - Built an app that allows users to find, search and host parties near you. Won this award in the Fall 2021 AppDev Hack Challenge at Cornell University.
- Columbia University Science Honors Program Scholar - Accepted by Columbia University to take undergraduate/graduate-level courses as part of their high school science honors program (approx. 10 % acceptance rate). Took Discrete Random Walks, Relativity and Quantum Physics, Bioinformatics, and Python Programming.
- Salutatorian of Valhalla High School

Skills and Activities

- Programming Languages: Python, Java, OCaml, C++, C, MATLAB, SQL
- Frameworks/Libraries: PyTorch, Tensorflow, NumPy, Hugging Face
- Hobbies: Badminton, Squash, Running