Artificial Intelligence Initiative (Aii)

https://ai.ucf.edu/



Invited Speaker

Dr. Aditya Desai, UC Berkeley

"Towards exponentially cheaper AI"

Thursday, August 29, 2024 • 10:00AM • HEC 101A • Zoom

Abstract:
The recent advancements in the capabilities of AI models have been extraordinary. However, training and deploying these models is prohibitively costly. The primary reason for increasing costs is the exponential increase in model sizes, which requires commensurate computing and memory resources. High resource usage in AI model training causes several issues: (1) Only a few large corporations can afford to train these models, limiting broader participation and growth in the AI community. (2) The process is costly and environmentally harmful. (3) The resulting models are often too large to be deployed on commonly used devices like phones. How do we make AI more resource-efficient? Existing research in efficiency can only provide a constant factor improvement in performance. Thus, to combat the exponential demand for resources, we need to rethink AI's efficiency more fundamentally. In this talk, I will discuss a new approach to making ML models efficient, drawing inspiration from probabilistic algorithms.

I will focus on our development of a new parameter memory approach called Elastic Memory (EM), which applies a probabilistic perspective to parameter memory. EM offers a superior memory-quality trade-off compared to traditional methods, supported by extensive empirical evidence and innovative theoretical analysis. We demonstrate that EM can exponentially reduce the size of deep learning recommendation models (from 100GB to 10MB) without compromising quality, achieving 3x faster latency and significantly lowering training and inference costs. EM can also be extended to enhance latency-quality trade-offs in ML, enabling 1.3x faster LLMs with consistent quality. I will conclude with a brief discussion on addressing deployment challenges in LLMs.

Bio:
Aditya recently completed his PhD at Rice University and joined UC Berkeley for postdoctoral research. His research focuses on evaluating the trade-offs between resources (such as memory, compute, and energy) and utility (including model quality metrics) in AI systems. His goal is to reduce resource usage while identifying representations and schemes that offer improved resource-utility trade-offs. He employs probabilistic algorithms and data structures to address scalability challenges in AI. Aditya's work was recognized at MLSys 2022 with an Outstanding Paper Award, and he received the Rice Future Faculty Fellowship in 2024. Additionally, he co-organized the Research on Algorithms & Data Structures (ROADS) to Mega-AI Models Workshop at MLSys 2023. Prior to his doctoral studies, Aditya spent five years as a Quantitative Analyst at Tower Research Capital after completing his undergraduate degree at IIT Kanpur.