

Speaker: Udit Gupta, Harvard University and Facebook

Title: Designing Specialized Systems for Deep Learning-based Personalized Recommendation

Abstract: We have witnessed a Cambrian explosion in AI technology over the past decade. The exponential rise is largely fueled by the design, development, and deployment of specialized software and hardware that have yielded orders of magnitude efficiency improvements for deep learning. Despite these efforts, this talk focuses on an important, yet under-studied area: systems for deep learning-based personalized recommendation. Personalized recommendations form the backbone of our interaction with the Internet from online search to e-commerce, from streaming services to social media. To provide personalized experiences, deep learning-based recommendation models consume the majority of training and inference cycles in production data centers. This talk highlights the unique challenges faced by deep learning models for production-scale recommendation compared to other AI applications. Given the unique challenges, we design solutions across the software and hardware stack to optimize inference efficiency by jointly considering models, data-center scale implications, and the underlying hardware. Finally, the talk charts directions for designing responsible recommender and AI systems --- particularly focusing on environmental sustainability.

Bio: Udit Gupta is a 6th year PhD student at Harvard University and visiting research scientist at Facebook AI Research. His research interests focus on improving the performance and efficiency of emerging applications in computer systems and architecture by co-designing solutions across the computing stack. His recent work focuses on the optimization of datacenter-scale deep learning-based personalized recommendation engines. Udit has co-founded the PeRSONAI (personalized recommendation systems and algorithms) workshop and CLEAR (computing landscapes with environmental accountability and responsibility) workshops co-located at systems and machine learning conferences. Udit is a steering committee member of the Computer Architecture Student Association. He graduated with a B.S. from Cornell University in 2016.