پژوهشگاه دانش‌های بنیادی
پژوهشکده علوم کامپیوتر

سخنرانی علمی

Reconciling Accuracy, Cost, and Latency of Inference Serving Systems

پویان جمشیدی، دانشگاه کارولینای جنوبی

Abstract

ML inference services serve user requests directly, requiring fast and accurate responses. Moreover, these services face dynamic workloads of requests, imposing changes in their computing resources, and failing to right-size computing resources results in either latency service level objectives (SLOs) violations or wasted computing resources. Adapting to dynamic workloads for ML inference pipelines is a difficult problem because of the exponentially large design space (multiple interacting and interdependent components, each with different knobs that influence performance) and multiple competing performance objectives and conflicting user preferences (accuracy, latency, and cost). In addition, the specific reconfiguration must be decided in real time and with incomplete and imperfect knowledge due to dynamic workload, variable network latency, and variable resource availability. In this talk, I will present our recent solutions to the abovementioned challenges: InfAdapter combines model-switching and auto-scaling to enable a more granular design space to trade accuracy, cost, and latency in inference serving systems; IPA dynamically reconfigures ML inference pipelines to achieve the tradeoff; and Sponge deals with dynamic SLOs and achieves its goal by applying in-place vertical scaling, dynamic batching, and request reordering.

Biography

Pooyan Jamshidi is an assistant professor in the University of South Carolina's Department of Computer Science and Engineering. Pooyan has also worked in the industry; most recently, he was a visiting researcher at Google in 2021. Before his current position, Pooyan was a postdoctoral researcher at Carnegie Mellon University (2016 - 2018) and Imperial College London (2014 - 2016). He received a Ph.D. in computer science from Dublin City University in 2014 and an M.S. and B.S. in Systems Engineering and Computer Science and Math from Amirkabir University of Technology in 2003 and 2006. Pooyan's research interests span the areas of Software, Systems, AI/ML, and Robotics. In particular, he is interested in developing algorithms and tools that enable building resilient systems that can automatically handle goal tradeoffs, incorporate user preferences and constraints, identify causes of failures, and self-adapt to be able to operate in dynamic environments. Pooyan's work integrates various areas such as distributed systems, control theory, statistical learning and optimization, causal inference, representation learning, and transfer learning, focusing on applications in autonomous systems, AI accelerators, and software/hardware co-design.

زمان: یکشنبه ۱۴۰۳/۰۴/۳۱ - ساعت ۱۴:۰۰

مکان: فرمانیه، خ لواسانی، نبش خ فربین، پژوهشگاه دانش‌های بنیادی، ط ۲، کلاس C

ارائه به صورت حضوری انجام خواهد شد.

***** شرکت برای عموم علاقه‌مندان آزاد است *****