A CASE FOR MODEL-LESS INFERENCE SERVING

Ph.D. candidate QIAN LI

Computer Science Department · Stanford University

Host：梁云 长聘副教授
2019年12月27日 星期五 3:00pm
理科一号楼 1126会议室

ABSTRACT: The number of applications relying on machine learning models is already large and expected to keep growing. Despite existing work in machine learning inference serving, ease-of-use and cost-efficiency remain key challenges. Developers must manually match the performance, accuracy, and cost constraints of their applications to decisions about selecting the right model and model optimizations, suitable hardware architectures, and auto-scaling configurations. Making these decisions is an error-prone and difficult task for users, especially when the application load varies, applications evolve, and the available resources vary over time. In this talk, I will (a) share a vision for model-less inference serving, (b) present our on-going work INFaaS, a model-less inference-as-a-service system, and (c) discuss open research directions.

BIOGRAPHY: She is a Computer Science Ph.D. candidate at Stanford University, advised by Professor Christos Kozyrakis. She has broad interests in computer systems and architecture. The way how hardware, software, and data interact with each other appeals to me the most. Her recent research focuses on building a machine learning serving system that enables ease-of-use and high efficiency. She is a member of the MAST research group and the Platform Lab at Stanford.She earned her M.S. in Computer Science at Stanford University in 2019. Prior to joining Stanford, she received my Bachelor of Science from School of EECS at Peking University in 2017. She graduated summa cum laude in Computer Science and Technology. She was a member of Center for Energy-efficient Computing and Applications (CECA) .