北京大学前沿计算研究中心
Center on Frontiers of Computing Studies, Peking University

静园5号院
青年讲座

Ekko: A Large-Scale Deep Learning Recommender System with Low-Latency Model Update

Dr. Luo Mai

Assistant Professor
University of Edinburgh

Host：董豪 助理教授
2022年7月12日 星期二 14:00-15:00
在线讲座

Abstract

Deep Learning Recommender Systems (DLRSs) need to update models at low latency, thus promptly serving new users and content. Existing DLRSs, however, fail to do so. They train/validate models offline and broadcast entire models to global inference clusters. They thus incur significant model update latency (e.g. dozens of minutes), which adversely affects Service-Level Objectives (SLOs).

In this talk, I will describe Ekko, a novel DLRS that enables low-latency model updates. Its design idea is to allow model updates to be immediately disseminated to all inference clusters, thus bypassing long-latency model checkpoint, validation and broadcast. To realise this idea, we first design an efficient peer-to-peer model update dissemination algorithm. This algorithm exploits the sparsity and temporal locality in updating DLRS models to improve the throughput and latency of updating models. Further, Ekko has a model update scheduler that can prioritise, over busy networks, the sending of model updates that can largely affect SLOs. Finally, Ekko has an inference model state manager which monitors the SLOs of inference models and rollbacks the models if SLO-detrimental biased updates are detected. Evaluation results show that Ekko is orders of magnitude faster than state-of-the-art DLRS systems. Ekko has been deployed in production for more than one year, serves over a billion users daily and reduces the model update latency compared to state-of-the-art systems from dozens of minutes to 2.4 seconds. Ekko has been accepted to OSDI 2022.

Biography

Luo Mai is an Assistant Professor in the School of Informatics at the University of Edinburgh. He does research at the intersection of computer systems, machine learning and data management. His research has led to publications in OSDI, NSDI, VLDB, USENIX ATC and CoNEXT. Before joining Edinburgh, Luo was a research associate at Imperial College London and a visiting researcher at Microsoft Research. Luo obtained his PhD from Imperial College London with the support of a Google PhD Fellowship.

http: //cfcs.pku.edu.cn/