Principled System Designs for Large-scale Machine Learning

Abstract:
At the core of machine learning (ML) is the interplay between model training and data representations. The performance of ML programs depends on both steps, and it is often a labor-intensive process to iterate between them in order to achieve desirable solutions. With the growing data and model complexity, ML development increasingly relies on large-scale systems. In this thesis we focus on addressing the challenges in scaling model training and data engineering to distributed setting by harnessing ML characteristics.

We propose principled computational frameworks that are general-purpose and user-programmable to support a broad range of ML applications. To scale out model training to distributed settings, we provide new theoretical analyses for bounded staleness consistency models and propose a new protocol that improves ML convergence stability and speed. Based on the proposed consistency, we create Bosen, a parameter server system that optimizes network usage via bandwidth-managed communication. By deciding what to send and when independently from the ML computation, Bosen utilizes all the available network bandwidth and communicates important updates early, achieving 3x convergence speedup on certain ML applications without any user code change. To scale data engineering, we propose a distributed data engineering framework that scales to billions of features and integrates with existing ML training systems, while providing a standardized library and a programmable interface for implementing custom data transformations. By considering data engineering and model training jointly, this thesis advances the system designs to achieve scalable ML toolchains.

Speaker:
Wei (David) Dai

Thesis Committee:
Eric Xing (Chair)
Gregory Ganger
Andy Pavlo
Joseph Gonzalez (UC Berkeley)

Dec. 1, 2017
1:30pm
8102 GHC

Link to draft document: