Carnegie Mellon University
Computer Science Department

THESIS ORAL

Aurick Qiao

Elastic Machine Learning Systems with Co-adaptation

Wednesday, August 4, 2021 – 12:00 p.m. – REMOTE

In recent years, the amount of computation being invested into machine learning (ML) and deep learning (DL) training has multiplied by several orders of magnitude. Under these conditions, elasticity (the ability of a system to dynamically adapt to changing supply and demand of compute resources over time) is a key ingredient for efficient resource management. Elasticity has long been proven to improve the resource utilization, execution performance, and fault tolerance of traditional applications such as web services and big data processing. However, elastic ML training is a relatively new area of interest, and faces different challenges from traditional applications due to ML training's highly sub-linear resource scalability, diverse execution patterns and strategies, and dependence between distributed workers.

This thesis steps beyond the existing early work in elastic ML by employing co-adaptation, i.e. combining both system-level and application-side adaptations, to better adapt to dynamic compute resources. Although previous frameworks can enable elasticity by relying on system-level implementations, they ignore the inherent resource adaptability of ML training that can be leveraged to better overcome the aforementioned challenges. We present the design, implementation, and evaluation of three elastic systems for ML that improve DL training time in shared GPU clusters by 37-50%, enable elasticity for a diverse set of ML training applications, and reduce the impact of resource failures by 78-95%.

Thesis Committee:
Eric P. Xing, Chair
Gregory R. Ganger
Phillip B. Gibbons
Joseph E. Gonzalez, UC Berkeley