TVM: An Automated Deep Learning Compiler

Tianqi Chen

12/08/2020

Abstract

Data, models, and computing are the three pillars that enable machine learning to solve real-world problems at scale. Making progress on these three domains requires not only disruptive algorithmic advances but also systems innovations that can continue to squeeze more efficiency out of modern hardware. Learning systems are in the center of every intelligent application nowadays. However, the ever-growing demand for applications and hardware specialization creates a huge engineering burden for these systems, most of which rely on heuristics or manual optimization. In this talk, I will present a new approach that uses machine learning and compilation to automate system optimizations. I will describe our approach in the context of deep learning deployment problems. I will first discuss how to design invariant representations that can lead to transferable statistical cost models, and apply these representations to optimize tensor programs used in deep learning applications. I will then describe the system improvements we made to enable diverse hardware backends. TVM, our end-to-end compiler, delivers performance across hardware back-ends that are competitive with state-of-the-art, hand-tuned deep learning frameworks.