JDS

Data Management for ML-based Analytics and Beyond

DANIEL KANG, Stanford University, USA
JOHN GUIBAS, Stanford University, USA
PETER BAILIS, Stanford University, USA
TATSUNORI HASHIMOTO, Stanford University, USA
YI SUN, University of Chicago, USA
MATEI ZAHARIA, Stanford University, USA

PROBLEM STATEMENT

This paper considers the problem of *end-to-end* machine learning (ML) deployments, from the collection of training data all the way to answering queries using ML models. We focus on errors in ML models and the data used to train them, along with algorithms for end-to-end uses of these models.

METHODS

We provide simple abstractions, model assertions and learned observation assertions (LOA) to find errors that are pervasive in ML model deployments and the data used to train these ML models. We further implemented our abstractions in open-source APIs. Our abstractions and APIs are easy for those who are not experts in ML to use and deploy.

RESULTS

We show that model assertions and LOA can be deployed in as few as 10 lines of code per assertion. They can find errors with up to 100% precision across domains ranging from video analytics, tabular data analytics, and translation.

SIGNIFICANCE

It is standard in the literature to assume that training data is "gold" (i.e., 100% accurate) and that bulk ML model metrics such as accuracy properly reflect model performance. We show that these are not the case, even in widely studied settings. Our simple abstractions point towards methods of checking end-to-end deployments. We hope that future work builds on our APIs.

Authors' addresses: Daniel Kang, Stanford University, Stanford, USA, ddkang@stanford.edu; John Guibas, Stanford University, Stanford, USA, jtguibas@stanford.edu; Peter Bailis, Stanford University, Stanford, USA, pbailis@cs.stanford.edu; Tatsunori Hashimoto, Stanford University, Stanford, USA; Yi Sun, University of Chicago, Chicago, USA; Matei Zaharia, Stanford University, Stanford, USA, matei@cs.stanford.edu.

Research Article