Designing Production-Friendly Machine Learning

Matei Zaharia
Stanford and Databricks
matei@cs.stanford.edu

ABSTRACT

Building production ML applications is difficult because of their resource cost and complex failure modes. I will discuss these challenges from two perspectives: the Stanford DAWN Lab and experience with large-scale commercial ML users at Databricks. I will then present two emerging ideas to help address these challenges. The first is "ML platforms", an emerging class of software systems that standardize the interfaces used in ML applications to make them easier to build and maintain. I will give a few examples, including the open-source MLflow system from Databricks [3]. The second idea is models that are more "production-friendly" by design. As a concrete example, I will discuss retrieval-based NLP models such as Stanford's ColBERT [1, 2] that query documents from an updateable corpus to perform tasks such as question-answering, which gives multiple practical advantages, including low computational cost, high interpretability, and very fast updates to the model's "knowledge". These models are an exciting alternative to large language models such as GPT-3.

PVLDB Reference Format:
Matei Zaharia. Designing Production-Friendly Machine Learning. PVLDB, 14(13): 3420-3420, 2021.
doi:10.14778/3484224.3484241

BIOGRAPHY

Matei Zaharia is an Assistant Professor of Computer Science at Stanford and Cofounder and Chief Technologist at Databricks. He started the Apache Spark open-source project during his PhD at UC Berkeley in 2009 and the MLflow open-source project at Databricks, and has helped design other widely used data and AI systems software including Delta Lake and Apache Mesos. At Stanford, he is a co-PI of the DAWN Lab working on infrastructure for machine learning, data management, and cloud computing. Matei's research was recognized through the 2014 ACM Doctoral Dissertation Award for the best PhD dissertation in computer science, an NSF CAREER Award, and the US Presidential Early Career Award for Scientists and Engineers (PECASE).

REFERENCES

[1] Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided Supervision for OpenQA with ColBERT. *Transactions of the Association for Computational Linguistics (TACL)* 9 (2021), 929–944.
[2] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *ACM International Conference on Research and Development in Information Retrieval (SIGIR)*. 39–48.
[3] Matei Zaharia, Andrew Chen, Aaron Davidson, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, Fen Xie, and Corey Zumar. 2018. Accelerating the Machine Learning Lifecycle with MLflow. *IEEE Data Engneering Bulletin* 41, 4 (2018), 39–45.

Proceedings of the VLDB Endowment, Vol. 14, No. 13 ISSN 2150-8097.
doi:10.14778/3484224.3484241