Data Attribution: A Data-Centric Approach for Trustworthy AI Development

Jiaqi Ma

University of Illinois Urbana-Champaign
jiaqima@illinois.edu

Abstract

Data plays an increasingly crucial role in both the performance and the safety of AI models. Data attribution is an emerging family of techniques aimed at quantifying the impact of individual training data points on a model trained on them, which has found data-centric applications such as instance-based explanation, unsafe training data detection, and copyright compensation. In this talk, I will comprehensively review our work contributing to the applications, methods, and open-source benchmarks of data attribution, and discuss open challenges in this field.