Countering Adversarial Attacks in Distributed Learning via User-Level Privacy

Proposed by :
Dionysis Manousakas
University of Cambridge

dm754@cam.ac.uk

Abstract

Lack of certified robustness to adversarial attacks is one of the major concerns for large scale deployment of learning in distributed settings. As originally proposed, learning via federated averaging algorithms is prune to targeted and untargeted attacks from involved parties, which can be achieved via malicious parameter updates communicated from distributed devices.

Differentially private mechanisms allow a composable framework for privacy protected access to a database containing sensitive information of individuals: Differential privacy primarily imposes insensitivity conditions at the output of a mechanism with regards to the removal of parts of the input dataset.

In this project we explore scalable robustness enhancing adjustments of existing distributed learning algorithms (such as gradient clipping and user privacy) and investigate their efficiency (in terms of accuracy and running time) in large scale deep learning applications.

Keywords
distributed learning, federated learning, asynchronous algorithms,
adversarial attacks, deep learning, certified robustness, differential privacy