Seminar/Talk

Adversarial Training Should Be Cast as a Non-Zero-Sum Game

Volkan Cevher
EPFL

Host: Marco Mondelli

One prominent approach toward resolving the adversarial vulnerability of deep neural networks is the two-player zero-sum paradigm of adversarial training, in which predictors are trained against adversarially-chosen perturbations of data. Despite the promise of this approach, algorithms based on this paradigm have not engendered sufficient levels of robustness, and suffer from pathological behavior like robust overfitting. To understand this shortcoming, we first show that the commonly used surrogate-based relaxation used in adversarial training algorithms voids all guarantees on the robustness of trained classifiers. The identification of this pitfall informs a novel non-zero-sum bilevel formulation of adversarial training, wherein each player optimizes a different objective function. Our formulation naturally yields a simple algorithmic framework that matches and in some cases outperforms state-of-the-art attacks, attains comparable levels of robustness to standard adversarial training algorithms, and does not suffer from robust overfitting.

Monday, March 25, 2024 01:00pm - 02:00pm

Office Bldg West / Ground floor / Heinzel Seminar Room (I21.EG.101)

This invitation is valid as a ticket for the ISTA Shuttle from and to Heiligenstadt Station. Please find a schedule of the ISTA Shuttle on our webpage: https://ista.ac.at/en/campus/how-to-get-here/ The ISTA Shuttle bus is marked ISTA Shuttle (#142) and has the Institute Logo printed on the side.