Project

Defending against model-stealing attacks targeting “Machine Learning as a Service” (MLaaS) platforms has proven extremely difficult. While various approaches, such as monitoring access to the model, perturbing/obfuscating results, and watermarking have been explored in the past, all have specific weaknesses, and there is no silver bullet to defense. In this project, we analyze existing approaches against model-stealing attacks, explore their limits, and set out to develop a new, more generic defensive mechanism.

Contact / Supervision

Yilin Ji, yilin.ji@kit.edu, Room 163 (Geb. 50.34)

References


