Karlsruher Institut für Technologie

Institut für Informationssicherheit und Verlässlichkeit (KASTEL)
Intelligent System Security
Jun.-Prof. Dr. Christian Wressnegger

Efficient Model Stealing Attacks

Project Group "Praxis der Forschung" – Winter Term 2021/22

Project

In recent years, "Machine Learning as a Service" (MLaaS) has become a cost-effective alternative to learning models on-site. Operators of such services charge users by the number of queries and, thus, take a great interest in keeping the underlying model private. However, by strategically querying inputs close to the decision boundary, it is possible to "steal" the model. This project systematically explores the prerequisites of recently proposed approaches, highlights their drawbacks, and develops a more query-efficient approach that does not rely on knowledge of the underlying data distribution.

Contact / Supervision

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

References

[1] U. Aïvodji, A. Bolot, and S. Gambs. Model extraction from counterfactual explanations. *CoRR*, abs/2009.01884, 2020.

[2] V. Chandrasekaran, K. Chaudhuri, I. Giacomelli, S. Jha, and S. Yan. Exploring connections between active learning and model extraction. In *Proc. of the USENIX Security Symposium*, pages 1309–1326, 2020.

[3] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Stealing machine learning models via prediction apis. In *Proc. of the USENIX Security Symposium*, pages 601–618, 2016.

[4] M. Yan, C. W. Fletcher, and J. Torrellas. Cache telepathy: Leveraging shared resource attacks to learn DNN architectures. In *Proc. of the USENIX Security Symposium*, pages 2003–2020, 2020.

[5] H. Yu, K. Yang, T. Zhang, Y. Tsai, T. Ho, and Y. Jin. Cloudleak: Large-scale deep learning models stealing through adversarial examples. In *Proc. of the Network and Distributed System Security Symposium (NDSS)*, 2020.