Karlsruher Institut für Technologie

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

Improving Defenses Against Model Stealing

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

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, perturbating/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

[1] M. Juuti, S. Szyller, S. Marchal, and N. Asokan. PRADA: protecting against DNN model stealing attacks. In *Proc. of the IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 512–527, 2019.

[2] M. Kesarwani, B. Mukhoty, V. Arya, and S. Mehta. Model extraction warning in MLaaS paradigm. In *Proc. of the Annual Computer Security Applications Conference (ACSAC)*, pages 371–380, 2018.

[3] T. Lee, B. Edwards, I. M. Molloy, and D. Su. Defending against neural network model stealing attacks using deceptive perturbations. In *Proc. of the IEEE Symposium on Security and Privacy Workshops*, pages 43–49, 2019.

[4] T. Orekondy, B. Schiele, and M. Fritz. Prediction poisoning: Towards defenses against DNN model stealing attacks. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2020.

[5] S. Pal, Y. Gupta, A. Shukla, A. Kanade, S. K. Shevade, and V. Ganapathy. Activethief: Model extraction using active learning and unannotated public data. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 865–872, 2020.

[6] 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.

[7] 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.

[8] 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.