Private Machine Learning: Effectiveness against Attacks

supervised by Aleksei Triastcyn

Keywords: deep learning, differential privacy

Description

Recent advances in machine learning, and especially deep learning, allow companies, governments and individuals to benefit from vast amounts of accumulated data. On the other hand, the ability of deep learning models to capture fine levels of detail can potentially compromise privacy of users. Recent research [1, 2] suggests that even in a black-box setting it is possible to detect the presence of individual records in the training set or recover certain features of these records. It necessitates research and development of the matching defences. One of the directions for such research is based on the widely accepted notion of *differential privacy* [3].

The goal of this project is to evaluate the resistance of machine learning models and appropriate defence mechanisms to privacy attacks (such as membership inference [2], model inversion [1], or dataset poisoning). The work will include reading relevant papers, implementing evaluation metrics and attack algorithms in Python, designing and running experiments.

References

[1] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1322–1333.

[2] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," In *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 2015, pp. 1322–1333.

[3] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.