Ownership Protection in Machine Learning Processes

Tanja Šarčević
SBA Research and *Vienna University of Technology*
Vienna, Austria
tsarcevic@sba-research.org
DOI: 0000-0003-0896-9193

***Abstract*—Outsourcing and shifting data storage and complex Machine Learning models to cloud services witnessed great growth over the past years, as costs of producing, maintaining and processing data can be decreased this way. However, sharing these assets entails potential intellectual property (IP) theft, and existing mechanisms for IP protection are susceptible to attacks. Our work develops methods for protecting IP privacy on multiple levels of the Machine Learning process, i.e. IP protection of shared input data and IP protection of derived Machine Learning models. The research results provide novel schemes for IP protection and better insights into the effects of these schemes on the quality and utility of the affected intellectual property assets.**

I. Introduction

Sharing digital assets is as old as digital data itself. Data creators are sharing data for commercial reasons, for increasing their reputation for valuable creations or information, for research reasons, to support the preservation of data long-term, etc. And over the last decades, the trend of sharing and processing digital data has vastly increased. Since data is a valuable asset to its owner, any type of unauthorised usage of shared data should be detected and sanctioned. With the advances in the area of Machine Learning (ML), outsourcing data and processing thereof became an increasingly popular trend among businesses. In this scope, the data is given to data management professionals that are involved in data mining, data classification etc., to make more use of the data more. This can foster business growth by additional services (recommendation systems) or customer behaviour understanding. In healthcare, for example, medical data are shared with researchers for help in medical diagnosis or other types of services that ML may provide. Furthermore, sharing the Machine Learning models has an increasing popularity through the online services such as Machine-Learning-as-a-Service (MLaaS). These platforms facilitate development, especially for researchers and small or medium businesses because in many ML settings, training an effective model from scratch requires a lot of computational power, human expertise and amount of data (e.g. natural language processing, image processing, etc.) In line with this, the model owners that have invested significant resources to train a model consider IP protection methods to verify the ownership or prevent unauthorised usage of a model.

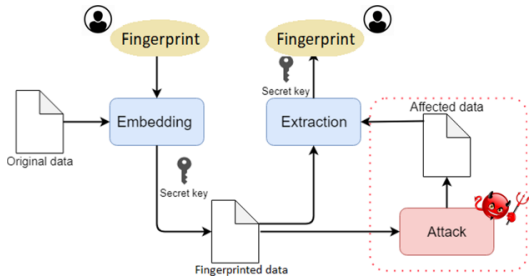

Fig. 1: Fingerprinting process (applicable to watermarking)

Digital watermarking and *fingerprinting* are approaches for protecting ownership of various types of digital property, including those relevant in ML process - data and ML models. By embedding a mark into a digital object these methods enable the owners to share these objects in their full form while enabling ownership claim and/or tracing recipients. Watermarking enables the owner of an object to verify their ownership, while fingerprinting, in addition, allows tracing the source of the unauthorised usage, i.e. the object-receiving party that re-shared the asset without the owner's authorisation. Hence, in case of fingerprinting, a unique mark is created and embedded in the object for every recipient.

There are multiple methods proposed for watermarking and fingerprinting data and ML models, generally adhering to a 2-step process shown in Figure 1: (i) watermark/fingerprint embedding (or insertion) and (ii) watermark/fingerprint detection (or extraction). The requirements for such process are:

1) recognisable by the owner
2) not detectable (and removable) by the recipients
3) robust to the modifications of the object
4) utility of the object is preserved

To achieve these 4 requirements simultaneously, the techniques need to ensure that the enough marks are embedded to achieve good robustness, but within certain limits to preserve the utility, hence achieving a trade-off between robustness and utility is one of the major challenges in this topic.

A. Related work

The first watermarking and fingerprinting techniques for data were developed for multimedia content [1] and later extended to other domains such as graphs [2], sequential

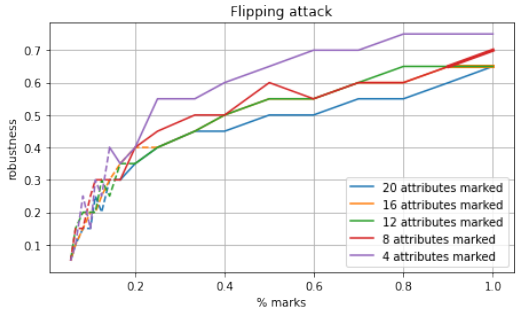

Fig. 2: Flipping attack on German Credit data[2]

data [3] and relational datasets [4], [5]. Our focus is mostly on the later. Relational data offers additional challenges compared to multimedia data for watermark/fingerprint embedding such as a limited redundancy of data representation for information (mark) hiding and non-uniformity of data content, i.e. a mix of data attribute types (numerical, text, categorical, etc.). Watermarking methods for ML models is a research topic that gained interest in the last 4 years and most of the techniques are designed specifically for Deep Neural Networks for image classification tasks. The methods are classified in two main groups depending on the scenario of watermark extraction. The *white-box* techniques allow extracting the marks from the model if it is available (shared without and authorisation) in its full form (model parameters, hyperparameters, etc.) [6], [7]. The *black-box* techniques allow extracting the watermark from the predictions of the model [8]–[10].

II. Preliminary research results

A. Fingerprinting relational data

In out work, we address a few shortcomings in the area of fingerprinting relational data. Firstly, most of the techniques are developed for numerical data in the relational data sets. We close this gap by proposing a technique applicable to categorical data, that preserves the original correlations in the data [11]. Secondly, due to the general lack of open-source implementation of fingerprinting techniques for relational data and the lack of conceptual methods for fingerprinting different data types, e.g., categorical, decimal, etc., we have adapted some of the existing techniques to provide the implementation of fingerprinting methodology that can be applied on a variety of data types within relational databases in the form of open-source fingerprinting toolbox[1]. The toolbox in addition contains implementations of existing fingerprinting methods and framework for simulation of the most common attacks against fingerprinting methods to enable robustness evaluation. Using this framework, we evaluated robustness of various techniques from the literature against the attack to obtain the understanding how the choice of the fingerprint parameters affects the resilience of the fingerprint and identify robust scenarios. Figure 2 shows such robustness results for our implemented fingerprinting scheme against an exemplary malicious attack, *flipping* the set of random values in the data set. In an addition to the robustness evaluation, the utility of the fingerprinted data was evaluated to address the trade-off between robustness and preserved data utility that needs to be found for successful fingerprint embedding. [12] The details on our contributions in the fingerprinting relational data and future challenges are summarised in [13].

[1]https://github.com/tanjascats/fingerprinting-toolbox, version 0.1.0

[2]https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

B. Watermarking machine learning models

The work is focusing on a several objectives in the area of watermarking ML models. One is the development and application of existing methods in order to assess their differences and shortcomings from the robustness point of view and utility of the fingerprinted or watermarked models. Another objective is focusing on specific attack towards and exploring new vulnerabilities of watermarking schemes. Furthermore, in the current research we focus on watermarking methods for special settings such as federated learning, where the existing methods cannot be directly applied.

References

[1] W. Trappe, M. Wu, Z. J. Wang, and K. R. Liu, "Anti-collusion Fingerprinting for Multimedia," *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1069–1087, 2003.

[2] M. Piec and A. Rauber, "Real-Time Screen Watermarking Using Overlaying Layer," in *Proceedings of the 9th International Conference on Availability, Reliability and Security*, 2014, pp. 561–570.

[3] E. Ayday, E. Yilmaz, and A. Yilmaz, "Robust {Optimization-Based} watermarking scheme for sequential data," in *22nd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2019)*, 2019, pp. 323–336.

[4] R. Agrawal and J. Kiernan, "Watermarking Relational Databases," in *Proceedings of the 28th International Conference on Very Large Databases*. Elsevier, 2002, pp. 155–166.

[5] Y. Li, V. Swarup, and S. Jajodia, "Fingerprinting Relational Databases: Schemes and Specialties," *IEEE Transactions on Dependable and Secure Computing*, vol. 2, no. 1, pp. 34–45, 2005.

[6] B. Darvish Rouhani, H. Chen, and F. Koushanfar, "Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks," in *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019, pp. 485–497.

[7] H. Chen, B. D. Rouhani, C. Fu, J. Zhao, and F. Koushanfar, "Deepmarks: A secure fingerprinting framework for digital rights management of deep learning models," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 2019, pp. 105–113.

[8] H. Chen, B. D. Rouhani, and F. Koushanfar, "Blackmarks: Black-box multibit watermarking for deep neural networks," *arXiv preprint arXiv:1904.00344*, 2019.

[9] Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 1615–1631.

[10] J. Zhang, Z. Gu, J. Jang, H. Wu, M. P. Stoecklin, H. Huang, and I. Molloy, "Protecting intellectual property of deep neural networks with watermarking," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, 2018, pp. 159–172.

[11] T. Sarcevic and R. Mayer, "A correlation-preserving fingerprinting technique for categorical data in relational databases," in *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, 2020, pp. 401–415.

[12] T. Šarčević and R. Mayer, "An Evaluation on Robustness and Utility of Fingerprinting Schemes," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Cham, Switzerland: Springer, 2019, pp. 209–228.

[13] T. Sarcevic, R. Mayer, and A. Rauber, "Fingerprinting relational data," 2021.