Prodomo, V.; González, R.; Gramaglia, M. Trading accuracy for privacy in machine learning tasks: an empirical analysis. In: *2021 IEEE Conference on Communications and Network Security (CNS), 4-6 Oct. 2021, Tempe, AZ, USA (Virtual conference). IEEE, 2022, 2 p.*

DOI: 10.1109/CNS53000.2021.9729036

Trading Accuracy for Privacy in Machine Learning Tasks: An Empirical Analysis

Vittorio Prodomo
University Carlos III of Madrid
Leganes, Spain
vprodomo@it.uc3m.es

Roberto Gonzalez
NEC Laboratories Europe
Heidelberg, Germany
roberto.gonzalez@neclab.eu

Marco Gramaglia
University Carlos III of Madrid
Leganes, Spain
mgramagl@it.uc3m.es

***Abstract*—Different kinds of user-generated data are increasingly used to tailor and optimize, through Machine Learning, the operation of online services and infrastructures. This typically requires sharing data among different partners, often including private data of individuals or business confidential data. While this poses privacy issues, the current state-of-the-art solutions either impose strong assumptions on the usage scenario or drastically reduce the data quality. In this paper, we evaluate through a generic framework the trade-offs between the accuracy of Machine Learning tasks and the achieved privacy (measured as similarity) on the input data, discussing trends and ways forward.**

***Index Terms*—Machine Learning, Privacy, Trade offs**

I. Introduction

The number of Machine Learning (ML) applications that are used in production real-world environments has rocketed in the past years, following the amazing advances obtained in different areas. These ML-based applications range from the personalization of services or the improved healthcare offered to final users to the automatic management of networks by Telco operators in the new 5G architectures. However, these applications rely on input data coming from possibly heterogeneous sources (either human or other machines), and spread through platforms owned by different actors which may not be fully trusted. Clearly, this poses different privacy and confidentiality issues. Thus, before being processed, data shall be conveniently transformed to obtain privacy-preserving properties.

In this work, we propose an empirical evaluation of how different data transformation methods targeting privacy preservation perform in practical scenarios. Designing a generic framework that can prevent any information leakage with little or no prior assumption is unsurprisingly very challenging. Among other categorizations [1], the state of the art solutions can be classified according to the part of the system that requires privacy protection, depending on whether they are targeting (i) the Machine Learning model itself, redesigning training algorithms and pipelines to have specific privacy-preserving properties; (ii) the output, pruning some information and protecting against over-fitting prevention techniques, which have been proven successful in black-boxing attacks; and (iii) the input data, pre-processing it before its disclosure to e.g. data brokers, to ensure the privacy of sample features and sometimes also statistical information about groups of samples.

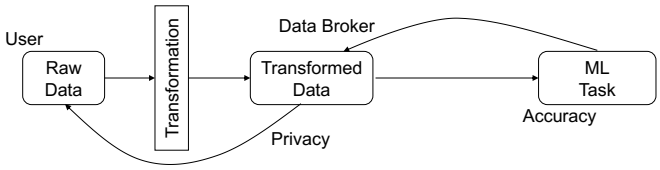

Fig. 1. The target scenario.

In this work, we focus on the latter approach. This family of techniques allows the sharing of training data among different actors without the privacy risks associated (i.e., the Privacy-Preserving Data Publishing, PPDP [2]). Thus, these technologies open the possibility of novel applications only possible before among trusted parties in (complex) federated learning scenarios.

In this scenario (depicted in Fig. 1) the user (or data owner) applies a transformation to the raw data. Then, the transformed data can be shared with other actors that may use it to train a ML model (possibly by adding the transformed data to other transformed data she owns). Hence, In this work we provide the following contributions: (i) a generic pipeline for measuring the similarity (which we use as an inverse proxy of privacy) between the transformed and raw data eventually used to perform a generic ML task (see Sec. II) and (ii) a discussion on the obtained results, showing the potential and the opportunities of the proposed methodology (see Sec. III).

II. Methodology

The goal of our methodology is to provide a generic pipeline to measure the accuracy versus privacy trade-off when generic ML tasks are performed on anonymized data, implementing the scenario depicted in Fig. 1. To showcase the generality of our pipeline we implement different transformation strategies, analyzing them in two ways: by measuring the similarity between the raw and the transformed data, which we use as a proxy of the achieved privacy level, and the accuracy of the performed machine learning task on the transformed data.

Data Transformation Strategies: In this work, we limit the study to a set of solutions that do not impose any specific knowledge on the input data and we set the basis for the study of other transformation algorithms. We first analyze noise addition, as proposed in [3], which advocates for randomly

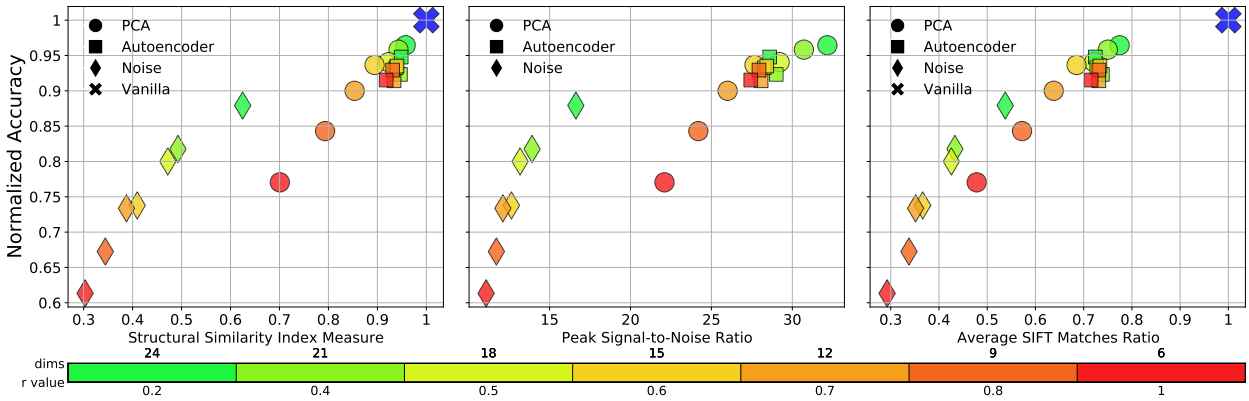

Fig. 2. The accuracy vs similarity tradeoffs for different similarity metrics. The accuracy results are normalized by the accuracy yielded by the raw data. Different markers represent the transformation strategies, while the color scale indicates their intensity (i.e., the amount of noise or the number of dimensions for the PCA and the AE).

(with a probability p) including White Gaussian Noise to the pixels that compose the image.

Then, we assess the performance of Principal Component Analysis (PCA), varying the number of retained Principal Components and reconstructing the data by reverting the operation starting from the retained dimensions. Finally, we evaluate Autoencoders (AE). We employ an 11-layer 2D Convolutional AE, trained for 100 epochs with a batch size of 256, a learning rate of 0.001, mean squared error as loss function, and Adam as optimizer. We revert to the original data shape following the same approach used for PCA.

Anonymization (Similarity) metric. Measuring privacy is complex, especially when no assumptions can be made on the kind of ML task, nor on the attacker models. To keep the system generic, we use Similarity measurements as a proxy for the achieved privacy levels. Thus, we resorted to numerical methods that measure the similarity between two items of data, and we chose three different methods to evaluate the similarity: (i) the Structural Similarity Index Measure (SSIM) [4], (ii) the Peak Signal-to-Noise ratio (PSNR), and (iii) a custom metric based on the Scale-invariant feature transform (SIFT) feature matching algorithm [5], which measure the ratio of common matching points between the original and transformed data.

Machine Learning algorithm. Finally, our pipeline should train a ML model to check the effect of the transformation on the accuracy. In this case, we use a *vanilla* image classification task, trained on the transformed data generated using one of the methods discussed above. We use a simple, 8-layer, 2D CNN to perform the classification task.

III. Experimental Evaluation

We test the framework depicted in Fig. 1 for an image classification task on the CIFAR-10 dataset [6]. Results are depicted in Fig. 2, showing the achieved trade-off between the two metrics we study, accuracy and privacy, the latter quantified with the methods presented in Sec. II.

For a system as the one targeted by this work, the ideal operational point lays in the top left corner: very high accuracy (comparable to the one achieved without any transformation) with very low similarity with the original data, hence maximizing the privacy level. The technique that, in general, better approximates that behaviour is the Noise addition one. By completely "hiding" features (pixels, in this case) this family of solutions achieves the lowest similarity level, still retaining enough accuracy for the envisioned Machine Learning task. The other solutions analyzed indeed better approximate the accuracy obtained by the not transformed data, but at a cost of a negligible loss in similarity, as most of the samples are in the top right corner for each of the discussed metrics.

IV. Conclusions

In this work, we presented an empirical analysis on the trade-offs that may be achieved in a data-sharing scenario where the machine learning task and the kind of attacker model are not known beforehand, a common scenario in a real operational environment. We tested several data transformation techniques, evaluated their accuracy, and discussed their effectiveness as privacy-preserving mechanisms.

V. Acknowledgement

The work of University Carlos III of Madrid was supported by the H2020 5G-TOURS project (grant no. 856950).

References

[1] I. Wagner and D. Eckhoff, "Technical privacy metrics: A systematic survey," *ACM Comput. Surv.*, vol. 51, no. 3, Jun. 2018.

[2] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, Jun. 2010.

[3] T. Zhang, Z. He, and R. B. Lee, "Privacy-preserving machine learning through data obfuscation," *CoRR*, vol. abs/1807.01860, 2018. [Online]. Available: http://arxiv.org/abs/1807.01860

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[5] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[6] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.