Macro-level Inference in Collaborative Learning[*]

Rudolf Mayer
rmayer@sba-research.org
SBA Research gGmbH
Vienna, Austria

Andreas Ekelhart
aekelhart@sba-research.org
SBA Research gGmbH
Vienna, Austria

ABSTRACT

With increasing data collection, also efforts to extract the underlying knowledge increase. Among these, collaborative learning efforts become more important, where multiple organisations want to jointly learn a common predictive model, e.g. to detect anomalies or learn how to improve a production process. Instead of learning only from their own data, a collaborative approach enables the participants to learn a more generalising model, also capable to predict settings not yet encountered by their own organisation, but some of the others. However, in many cases, the participants would not want to directly share and disclose their data, for regulatory reasons, or because the data constitute a business asset.

Approaches such as federated learning allow to train a collaborative model without exposing the data itself. However, federated learning still requires exchanging intermediate models from each participant. Information that can be inferred from these models is thus a concern. Threats to individual data points and defences have been studied e.g. in membership inference attacks. However, we argue that in many use cases, also global properties are of interest – not only to outsiders, but specifically also to the other participants, which might be competitors. In a production process, e.g. knowing which types of steps a company performs frequently, or obtaining information on quantities of a specific product or material a company processes, could reveal business secrets, without needing to know details of individual data points.

CCS CONCEPTS

• **Security and privacy → Distributed systems security**; • **Computing methodologies → Supervised learning**.

KEYWORDS

Federated learning, Adversarial attacks, Property inference

ACM Reference Format:
Rudolf Mayer and Andreas Ekelhart. 2022. Macro-level Inference in Collaborative Learning. In *Proceedings of the Twelveth ACM Conference on Data and Application Security and Privacy (CODASPY '22), April 24–27, 2022, Baltimore, MD, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3508398.3519361

1 INTRODUCTION

Collaborative learning has applications in many settings, e.g. in health care, where common models shall be obtained from distributed medical records. In this paper, we specifically focus on a setting where companies collaborate to obtain a shared model. This could be for anomaly detection of incidents such as fraud, where multiple credit-card providers want to collaborate to learn from fraudulent behaviour that yet has been observed only at a subset of the participants. Other application scenarios could e.g. be in manufacturing, where an automated assembly line is controlled by machine learning (ML) models, or robots in logistics, and collaborating manufacturers would like to learn how to perform certain, for them yet unknown procedures. While such common goals seems plausible, participants might not be willing to directly exchange their data due to privacy or confidentiality reasons.

Approaches such as federated learning alleviate many confidentiality issues of data sharing, by only exchanging an abstraction thereof, namely intermediate or final versions of a learned model, while providing comparable effectiveness. However, federated learning still poses risks, as the models itself can be attacked; due to its distributed nature, it might even open up novel attack vectors, compared to a centralised setting. Several types of attacks on ML pipelines have recently been studied. In terms of data confidentiality, emphasis is put on protecting micro-level, individual data, e.g. against attacks such as membership inference, using techniques such as differential privacy (or other noise-addition techniques) on the learned models. However, in this setting we consider use cases where the confidentiality of a different kind of information is at stake. Instead of needing to protect individual records, global properties of the data are our concern. Inferring e.g. from a competitor's model what kind of and which quantities of products they are producing might be more revealing than getting to know single data points. This setting has not yet received much attention in research, hence this paper introduces a threat model and highlights the need for solutions to protect this type of information to enable successful and mutually beneficial collaborative learning.

2 COLLABORATIVE ML THREAT MODEL

In this section, we discuss the goal and knowledge of attackers targeting machine learning (ML) systems, based on [3], with a focus on inference attacks on training data in **collaborative learning**.

Attacker's Goal: The **security violation** can be categorised along the axes of the so-called CIA triangle, which comprises attacks against integrity, availability, and confidentiality. Our primary concern is the confidentiality of data (and meta-data and meta-information) of the participants in collaborative learning. The **attack specificity** can be either targeted, i.e. inferring information on a specific data item or property, or untargeted, which entails revealing information on any kind of item or property.

[*] SBA Research (SBA-K1) is a COMET Centre within the framework of COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna; COMET is managed by FFG.
This work is supported by BMK and FFG under Grant No. 873979 (project PRIMAL).

CODASPY '22, April 24–27, 2022, Baltimore, MD, USA

ACM ISBN 978-1-4503-9220-4/22/04.
https://doi.org/10.1145/3508398.3519361

Attacker's Knowledge: Attacks can be distinguished based on the adversary's knowledge of the targeted system, divided into the following categories [7]: **Data knowledge** denotes information on the dataset and its distribution. An adversary often knows at least approximately how the training set is distributed and is therefore able to acquire a dataset with a similar distribution (*"shadow data"*). The original and shadow data are often disjoint, but might also be (partially) joint, if the attacker has sufficient information on the data set. **Model knowledge** denotes knowledge about the architecture and parameters of the trained model, e.g. the type of neuronal network, the activation functions and number of layers, and the learned parameters themselves. **Training knowledge** denotes knowledge on the learning algorithm, e.g. how the model was trained (the optimiser, number of epochs, other hyperparameters). **Output knowledge** is knowledge of the predictions, e.g. the class probability vector in a multi-class setting.

Based on the knowledge, we can then distinguish several scenarios. **White-box**: the attacker is assumed to know everything about the targeted system. **Black-box**: the attacker can only use the system, e.g. query inputs and observe outputs. **Gray-box**: any setting in between, i.e. the attacker knows parts of the system.

One can further distinguish different types of attackers in collaborative learning, depending on their role in the training process. An **insider attacker** participates in the process and has access to the models during training. We distinguish between a **participant** insider attacker who trains models locally [2], and an **aggregator** attacker that collects locally trained models [12]. An **outsider attacker** has access only to the final model after the collaborative learning process is finished, which is roughly equivalent to attacks on models trained in a centralised manner.

For many settings, one can further distinguish attackers based on their adherence to the collaboration protocol. **Semi-honest (or honest-but-curious)** adversaries perform a "passive" attack, i.e. they follow the protocol, but try to gather more information than it allows, e.g. information about the training data. **Malicious adversaries** perform "active" attacks and arbitrarily deviate from the protocol, e.g. with the goal to corrupt the learning process.

3 CONFIDENTIALITY ATTACKS AGAINST ML

Inference in the context of data publishing often distinguishes the following types of disclosure (e.g. [8]): *Identify disclosure*, where an attacker can match a specific record to an individual, *attribute disclosure*, where for an incomplete record of an individual the unknown values are inferred, and *membership inference*, where it is revealed whether a person was included in a dataset.

Releasing a trained model might as well cause unintentional information leakage about the training data, similar to leakage from published data. Frequently, the following attacks are discussed:

Model inversion is arguably the most powerful confidentiality attack against a model, and tries to re-create training data. The idea is that since a learned model stores a mapping between the input and output space, it can not only be used to infer predictions one way (i.e. from an input sample to an output), but may also be inverted to yield an optimal input so that the discrepancy between the predicted value and the target response (e.g. a specific class label) is minimised. This can be achieved by gradient descent methods, computing the local value of a loss function and incrementally approaching the most "correct" input. Fredrikson et al. [4] show this attack for a large feature space (an image with floating point pixel values) and re-create training data for logistic regression and (simple) neural networks for face recognition. A model is intentionally trained to generalise relevant class-inherent features – and thus, the re-created input will in most cases not represent a specific sample from the original data, but rather a form of average of features with the highest influence on the model's decision. This explains why model inversion is feasible in specific settings (e.g. face recognition, where one class resembles one individual), but not others (e.g. tasks where a class represents all members of a gender), and why even in face recognition, the results may rather resemble an unnatural caricature rather than a plausible photograph. Others have therefore tried to constrain the reconstructed data, by using e.g. generative adversarial networks trained on public data [13]; their results show that this improves the appearance by far, though it does not necessarily lead to images closer to the original data.

Model inversion pose few requirements on adversarial knowledge. Primarily, output knowledge is essential, as well as model knowledge (e.g. to perform gradient descent for input optimisation).

Attribute disclosure (or inference) can be seen as a special case of model inversion, requiring to invert only one (or a few) attributes from an otherwise known input sample. Fredrikson et al. [5] recover the genetic markers of individual patients from a pharmacogenetic linear regression model trained to predict drug dosing for patients based on their clinical history, demographics and genotype. This is achieved by a rather broad search over all possible value combinations, and eventually selecting those that produce the highest confidence. They also investigate the protective capacity of differential privacy against their attack, and show that it is not feasible – for differential privacy to reduce the attack success rate substantially, the model would lose its predictive power to an extent unacceptable in a clinical environment. Fredrikson et al. [4] later indicate that an adversary with white box access to a decision tree model can predict a sensitive feature with perfect precision.

Attribute disclosure requires more adversarial knowledge than model inversion. On top of output and model knowledge, at least for the data to be disclosed, partial inputs are needed.

Membership disclosure (or inference) assumes that the prediction (e.g. the vector of class-likelihoods) will exhibit distinguished patterns if they were part of the training data than if not – which is closely related to overfitting. The first work by [11] describes a supervised attack, and assumes that similar models trained on similar data must behave in a similar manner. Thus, the attacker tries to create such models (called "shadow models") from data assumed to be similar to the original. Based on the outputs for selected samples, and the information whether they were in the shadow training set, a so-called attack model is trained. Unsupervised attacks rely e.g. on the correctness of the prediction, or the prediction loss [10].

Membership disclosure requires significant adversarial knowledge. Output knowledge is needed to categorise inputs as members or not. For supervised settings, all of data, training and model knowledge are required to create the shadow models. Unsupervised attacks can suffice e.g. without knowledge of the model architecture and data distribution.

Property inference allows an attacker to extract properties of the dataset, even if they were not explicitly encoded as features or are not correlated to the learning task. [1] are among the first to show that it is possible to extract some characteristics of the training set that the effectiveness of the classifier might depend on, such as the prevalent accents in voice samples used to train a speech recognition software. A similar technique is used in [6] to infer training set properties from fully connected neural networks. In particular, a meta-classifier is trained on proxy models with the same task as the target model, while the training data was explicitly designed to have or not have the global or class-related target property. As an example, they try to infer if a specific dataset contained more male than female individuals.

4 MACRO-LEVEL THREATS IN COMMERCIAL COLLABORATIVE LEARNING

In the macro-level inference scenario of collaborative learning, we mostly consider a **white-box** setting, and attackers have all types of knowledge mentioned above – except the training data itself, as this is at least somewhat different for each participant (otherwise there would be a reduced incentive to collaborate). However, an attacker with a sufficiently similar dataset and domain knowledge available has approximate information on e.g. the possible value ranges of attributes, or the correlations between them.

Participants in the collaborative learning can be seen as potential **insider attackers**. They might be able to obtain the intermediate models of other participants, e.g. when there are peer-to-peer learning architectures employed. In other settings, they might infer information from the global model – maybe not in detail about specific participants, but about the union of all of them.

Attacks on the integrity and availability of the collaborative process are plausible, e.g. when a participant expects others to use the collaborative model and aims to disrupt its functionality. However, we primarily consider **confidentiality attacks**. The attack can be **targeted or untargeted** – in both settings, the attacker might learn valuable information to optimise its own business operation.

In terms of specific attacks, current research on property inference attacks indicates this attack is capable to disclose certain properties from a dataset. However, the properties inferred e.g. in [1] are still rather high-level, e.g. they infer that there is a higher proportion of a gender in the dataset, which is still very coarse; transferred to a commercial setting, knowing that a certain product is more prominent in a competitors operation than another product is maybe not detailed enough to capitalise on it, and more fine-grained information, e.g. the ratio, would be required. Thus, more research on the attack potential to uncover fine-grained details is required. Potentially, repeated property inference attacks could be carried out, each targeting a binary decision for a certain level of the potential range. Eventually, the most plausible ratio is selected as property value. The scalability of applying the meta-classifier based attack repeatedly also needs to be investigated.

While model inversion attacks are often discussed as targeting individuals represented in the data, a form of model inversion could also be a valid strategy in the macro-level setting. On the one hand, if it would be possible to re-create the training data set, then detailed information at least on representatives of different classes could be obtained. However, one can also envision settings where the attacker tries to estimate which types of well-understood classes (e.g. materials) the competitors have frequently in their training dataset, measured by the success of the inversion of those classes.

Furthermore, the transferability of relevant attacks to federated settings is yet not fully explored. While initial studies have been performed e.g. for membership inference (e.g. [9]), for property inference and model inversion, this still needs further analysis. Potential attack vectors include intercepting the local models from individual participants (e.g. in a sequential or peer-to-peer collaboration setting, or as a coordinator), or inferring information on local data from the global model that is shared back to participants.

5 CONCLUSION AND FUTURE WORK

Inference on global properties of data used in collaborative learning environments is still an unexplored topic, and further research is needed to fully understand the extent, severity and feasibility of such attacks. Based on such an analysis, fitting defence methods need to be developed. These likely need to differ from currently dominating approaches such as differential privacy, which protect data at the micro-scale.

REFERENCES

[1] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. 2015. Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers. *International Journal of Security and Networks* 10, 3 (2015).

[2] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2020. How To Backdoor Federated Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)* (Palermo, Italy).

[3] Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition* 84 (2018).

[4] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)* (Denver, USA). ACM.

[5] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In *USENIX Security Symposium* (San Diego, CA). USENIX Association.

[6] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. 2018. Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)* (Toronto, Canada). ACM.

[7] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2021. Membership Inference Attacks on Machine Learning: A Survey. arXiv: 2103.07853.

[8] Fabian Prasser, Florian Kohlmayer, Ronald Lautenschläger, and Klaus A. Kuhn. 2014. ARX–A Comprehensive Tool for Anonymizing Biomedical Data. *AMIA Annual Symposium Proceedings.*

[9] Anastassiya Pustozerova and Rudolf Mayer. 2020. Information Leaks in Federated Learning. In *Workshop on Decentralized IoT Systems and Security* (San Diego, CA). Internet Society.

[10] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Network and Distributed System Security Symposium* (San Diego, USA). Internet Society.

[11] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)* (San Jose, USA). IEEE.

[12] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. 2019. Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning. In *IEEE Conference on Computer Communications (INFOCOM)* (Paris, France). IEEE.

[13] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. 2020. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.