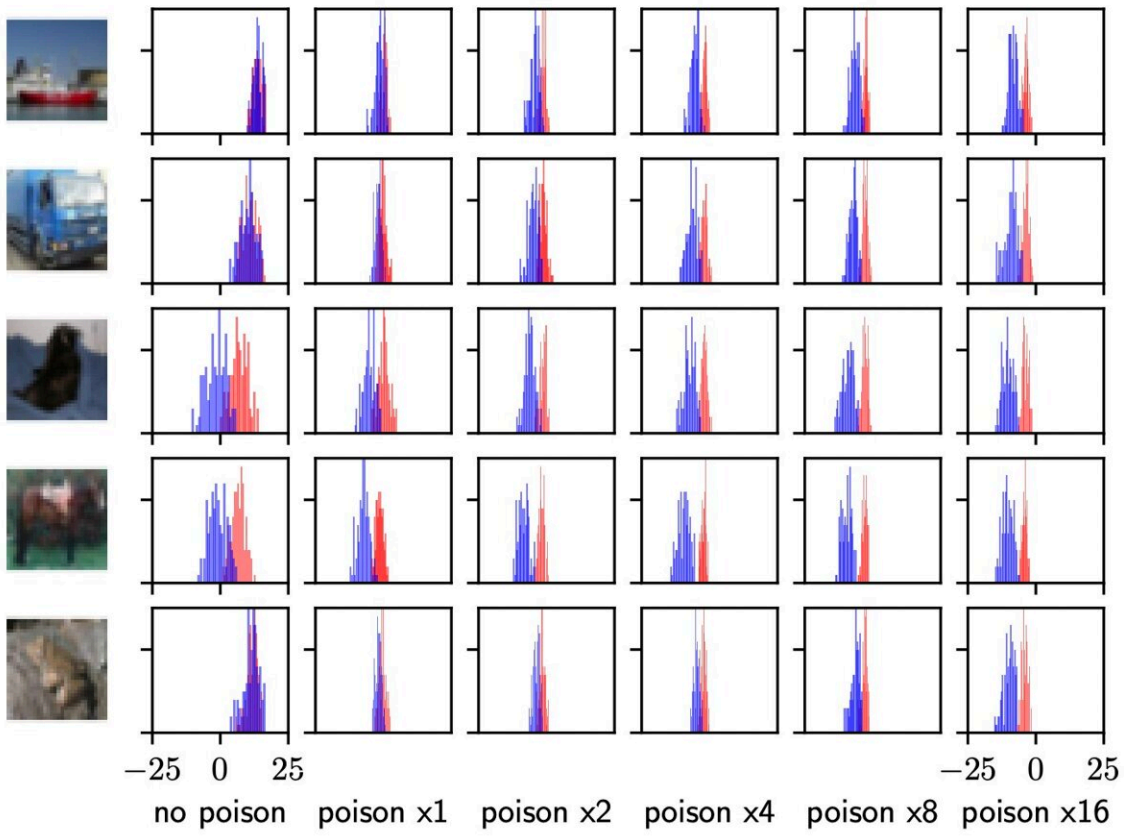
The risks of attacks that involve poisoning training data for machine learning models

April 25 2022, by Ingrid Fadelli



The attacks examined by the researchers separate the loss distributions of members and non-members, making them more distinguishable. For five random CIFAR-10 examples, this graph plots the (logit-scaled) loss distribution on one example, when it is a member (red) or not (blue). The horizontal axis varies based on the number of times the adversary poisons the example. Credit: Tramèr et al

A growing number of studies suggest that machine learning algorithms can leak a considerable amount of information included in the data used to train them through their model parameters and predictions. As a result, malicious users with general access to the algorithm can in many cases reconstruct and infer sensitive information included in the training dataset, ranging from simple demographic data to bank account numbers.

Researchers at Google, National University of Singapore, Yale-NUS College, and Oregon State University have recently carried out a study evaluating the risks of these type of attacks, which essentially entail "poisoning" machine learning models to reconstruct the sensitive information hidden within their parameters or predictions. Their paper, pre-published on arXiv, highlights the alarming nature of these attacks and their ability to bypass existing cryptographic privacy tools.

"The foundation of the adversary method is an inference algorithm, known as membership inference attack, that determines the chance that any arbitrary record has been part of the training set," Reza Shokri, one of the researchers who carried out the study, told TechXplore. "Inference attacks against ML is a serious data privacy threat because the adversary is a legitimate 'user' of the machine learning system and does not need to break into any system to gain access to sensitive information."

Previous studies by co-authors of the recent paper, as well as other research teams worldwide, have reported the privacy vulnerabilities of machine learning algorithms used in different settings, including ML-as-a-service platforms, federated learning tools, and large language models. In most of the attacks identified in these previous papers, excluding those involving federated learning settings, an adversary or malicious

user can perform inference attacks while merely "observing" the outcome of the learning process (i.e., the labels predicted by the model), yet he/she cannot influence the training process.

In their recent paper, Shokri and his colleagues specifically focused on the implementation of machine learning algorithms in a secure multi-party setting. In these cases, a model is trained on a combination of data that is independently provided by different individuals, developers, or other parties.

"Based on previous work in the field, we knew that the final model would leak some information about contributed training data by all parties," Shokri explained. "However, what we are showing in this paper is that a malicious party can significantly 'increase' the information leakage about other parties' data, by contributing adversarial data and poisoning the pool of training data."

Essentially, Shokri and his colleagues showed that, by "poisoning" the training data, a malicious user can prompt a training algorithm to "memorize" data provided by other parties. This in turn allows him/her to reconstruct their victim's data using a series of inference attacks. Inference attacks are data mining techniques that allow users to illegitimately gain knowledge about a person or company within a database.

In their paper, the researchers specifically evaluated the effectiveness and threat level of three different types of inference attacks, combined with the 'poisoning' of training data. They first looked at membership inference attacks, which allow attackers to determine whether a particular data record was part of the training dataset or not.

"The reason why these attacks are important is that they allow us to quantify how much information the models leak about the individual

data records in their training set," Shokri said. "Membership inference attacks are used to audit privacy in machine learning (e.g., tools such as ML Privacy Meter)."

In addition to membership inference attacks, Shokri and his colleagues evaluated the effectiveness of reconstruction attacks and attribute inference attacks. Both these attack subtypes allow adversaries to partially reconstruct the training data.

"For example, these attacks can allow users to generate sentences that significantly overlap with sentences used for training a language model or complete a sentence that, for instance start with Aleph One's credit card number is xxxxx, or infer a missing attribute about a partially known record (e.g., inferring the marital status of Aleph One)," Shokri said. "These attacks are usually based on membership inference attacks (i.e., membership inference attacks are used as a steppingstone to run reconstruction attacks)."

Shokri and his colleagues found that all the inference attacks they examined were alarmingly successful in the scenario they focused on, in which a user can poison a common pool of training data compiled by different users. This suggests that existing cryptographic privacy tools might not be enough to guarantee the privacy of users providing data to train machine learning algorithms.

"What we show, which is of a significant issue, is that the data points that are on average not leaked through regular inference attacks (without poisoning) become orders of magnitudes more vulnerable when an adversary is allowed to poison the training set," Shokri added. "Our results cast serious doubts on the relevance of cryptographic privacy guarantees in multiparty computation protocols for machine learning. We are now working towards designing powerful inference attacks to be able to provide an accurate privacy auditing for machine learning."

More information: Florian Tramèr et al, Truth serum: poisoning machine learning models to reveal their secrets. arXiv:2204.00032v1 [cs.CR], arxiv.org/abs/2204.00032

Reza Shokri et al, Membership inference attacks against machine learning models. arXiv:1610.05820v2 [cs.CR], arxiv.org/abs/1610.05820

Milad Nasr et al, Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. arXiv:1812.00910v2 [stat.ML], arxiv.org/abs/1812.00910

Nicholas Carlini et al, Extracting training data from large language models. arXiv:2012.07805v2 [cs.CR], arxiv.org/abs/2012.07805

Christopher A. Choquette-Choo et al, Label-only membership inference attacks. arXiv:2007.14321v3 [cs.CR], arxiv.org/abs/2007.14321

Citation: The risks of attacks that involve poisoning training data for machine learning models (2022, April 25) retrieved 27 January 2023 from https://techxplore.com/news/2022-04-involve-poisoning-machine.html