Poster:Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning

Ahmed Salem*, Apratim Bhattacharya†, Michael Backes*, Mario Fritz*, Yang Zhang*
*CISPA Helmholtz Center for Information Security
†Max Planck Institute for Informatics

Machine learning (ML) has progressed rapidly during the past decade. Nowadays, it has become the core component in many industrial domains ranging from automotive manufacturing to financial services. Leading Internet companies, such as Google,[1] Amazon,[2] and Microsoft,[3] further provide Machine Learning as a Service (MLaaS) to simplify ML deployment. In this setting, an MLaaS provider trains a machine learning model at their backend and provides the trained model to public as a black-box API.

The major factor that drives the current ML development is the unprecedented large-scale data. In consequence, collecting high-quality data becomes essential for building advanced ML models. Data collection is a continuous process as enormous data is being generated at every second. This turns ML model training into a continuous process as well: Instead of training an ML model for once and keeping on using it afterwards, the model provider, such as an MLaaS provider, needs to keep on updating the model with newly-collected data. In practice, this is also known as *online learning*. And we refer to the dataset used to perform model update as the *updating set*.

Regularly updating an ML model results in the model having different versions with respect to different model parameters. This indicates that if an ML model is queried with the same set of data samples at two different points in time, it will provide different outputs.

Our Contributions In this work, our main research question is: *Can different outputs of an ML model's two versions queried with the same set of data samples leak information of the corresponding updating set?*. This constitutes a new attack surface against machine learning models. Information leakage of the updating set can severely damage the intellectual property and data privacy of the model provider/owner.

We concentrate on the most common ML application – classification. More importantly, we target on black-box ML models – the most difficult attack setting where an adversary does not have access to her target model's parameters but can only query the model with her data samples and obtain the corresponding prediction results, i.e., *posteriors* in the case of classification.

In total, we propose four different attacks in this surface which can be categorized into two classes, namely, *single-sample attack class* and *multi-sample attack class*. The two attacks in the single-sample attack class concentrate on a simplified case when the target ML model is updated with one single data sample. We investigate this case to show whether an ML model's two versions' different outputs indeed constitute a valid attack surface. The two attacks in the multi-sample attack class tackle a more general and complex case when the updating set contains multiple data samples.

Among our four attacks, two (one for each attack class) aim at reconstructing the updating set which to our knowledge, are the first attempt in this direction. Compared to many previous attacks inferring certain properties of a target model's training set [1], dataset reconstruction attack leads to more severe consequences [3]. In theory, membership inference attacks [2], [4], [5] can also be leveraged to reconstruct the dataset from a black-box ML model. However, membership inference is not scalable in the real-world setting as the adversary needs to collect a large data sample which happens to include all the training set samples of the target model. Though our two reconstruction attacks are designed specifically for the online learning setting, we believe they can provide further insights on reconstructing a black-box ML model's training set in other settings.

Extensive experiments show that indeed, the output difference of the same ML model's two different versions can be exploited to infer information about the updating set.

General Attack Construction. Our four attacks follow a general structure, which can be formulated into an encoder-decoder style. The encoder realized by a multilayer perceptron (MLP) takes the difference of the target ML model's outputs, namely *posterior difference*, as its input while the decoder produces different types of information about the updating set with respect to different attacks.

To obtain the posterior difference, we randomly select a fixed set of data samples, referred to as the *probing set*, and probe the target model's two different versions (the second-version model is obtained by updating the first-version model with an updating set). Then, we calculate the difference between the two sets of posteriors as the input for our attack's encoder.

Single-sample Attack Class. The single-sample attack class contains two attacks: *Single-sample label inference attack* and *single-sample reconstruction attack*. The first attack predicts the label of the single sample used to update the target model.

[1] https://cloud.google.com/ml-engine/
[2] https://aws.amazon.com/machine-learning/
[3] https://azure.microsoft.com/en-us/services/machine-learning-studio/

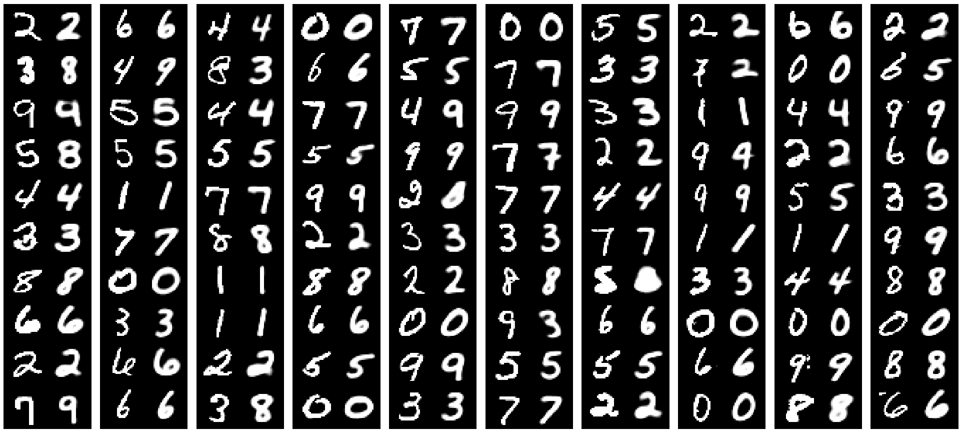

Fig. 1: Visualization of a full MNIST updating set together with the output of the multiple-sample reconstruction attack ($\mathcal{A}_{MSR}$) after clustering. The left column shows the original samples and the right column shows the reconstructed samples. The match between the original and reconstructed samples is performed by the Hungarian algorithm.

We realize the corresponding decoder for the attack by a two-layer MLP. Our evaluation shows that our attack is able to achieve a strong performance, e.g., 0.96 accuracy on the CIFAR-10 dataset.[4]

The single-sample reconstruction attack aims at reconstructing the updating sample. We rely on autoencoder (AE). In detail, we first train an AE on a different set of data samples. Then, we transfer the AE's decoder into our attack model as its sample reconstructor. Experimental results show that we can construct the single sample with a mean squared error (MSE) of 0.06355 for the MNIST dataset[5] and 0.01352 for the CIFAR-10 dataset, respectively. Moreover, we show that our attack learns to generate the specific sample used in the updating set [3], [5] instead of a general representation of samples affiliated with the same label.

Multi-sample Attack Class. The multi-sample attack class includes *multi-sample label distribution estimation attack* and *multiple-sample reconstruction attack*. Multi-sample label distribution estimation attack estimates the label distribution of the updating set's data samples. It is a generalization of the label inference attack in the single-sample attack class. We realize this attack by setting up the attack model's decoder as a multilayer perceptron with a fully connected layer and a softmax layer. Kullback-Leibler divergence (KL-divergence) is adopted as the model's loss function. Extensive experiments demonstrate the effecitiveness of this attack. For the CIFAR-10 dataset, when the updating set's cardinality is 100, our attack model achieves a 0.00376 KL-divergence which outperforms the baseline model by a factor of 3. Moreover, the accuracy of predicting the most frequent label is 0.32 which is also 3 times higher than the baseline model.

Our last attack, namely multiple-sample reconstruction attack, aims at generating all samples in the updating set. This is a much more complex attack than the previous ones. The decoder for this attack is assembled with two components. The first one learns the data distribution of the updating set samples. To this end, we propose a novel hybrid generative model, namely BM-GAN. Different from the standard generative adversarial networks (GANs), our BM-GAN introduces a "Best Match" loss which ensures that each sample in the updating set is reconstructed. The second component of our decoder relies on machine learning clustering to group the generated data samples by BM-GAN into clusters and take the central sample of each cluster as one final reconstructed sample. Our evaluation shows that we are able to reconstruct very similar samples as those in the original updating set on both MNIST and CIFAR-10 datasets. Figure 1 shows the final result for the multiple-sample reconstruction attack attack for the MNIST dataset, with an updating set of size 100.

To the best of our knowledge, this constitutes the first attack of this type, which is able to infer very detailed information on the dataset and even lends itself to full reconstruction of the data.

References

[1] M. Fredrikson, S. Jha, and T. Ristenpart, "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures," in *Proceedings of the 2015 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2015, pp. 1322–1333. 1

[2] Y. Long, V. Bindschaedler, L. Wang, D. Bu, X. Wang, H. Tang, C. A. Gunter, and K. Chen, "Understanding Membership Inferences on Well-Generalized Learning Models," *CoRR abs/1802.04889*, 2018. 1

[3] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Exploiting Unintended Feature Leakage in Collaborative Learning," in *Proceedings of the 2019 IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2019. 1, 2

[4] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models," in *Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019. 1

[5] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2017, pp. 3–18. 1, 2

[4] https://www.cs.toronto.edu/~kriz/cifar.html

[5] http://yann.lecun.com/exdb/mnist/