Summary Report: Machine Learning Models that Remember Too Much [1]

Koushik Chowdhury
Mat. Num: 2572865
MSc. Student, Saarland University

Data Privacy Seminar, CISPA, Saarland

Abstract

The main goal of the machine learning algorithm is to use train datasets to perform a variety of complex tasks and improve with the experience. We don't have to be an expert in machine learning to train datasets. There are lots of frameworks and services available nowadays. We have to take into account that when applying models to our datasets, no information about the training data should be lost. However, if we blindly use machine learning models on our datasets, we cannot stop the loss of information unless we are an expert in machine learning. The aim of this paper is to show that even if the training algorithm produces a model that is of high quality according to standard ML metrics such as accuracy, the adversary can still extract information from the ML model [1]. Throughout the experiment, datasets such as CIFAR10, LFW, FaceScrub, Newsgroup data and IMDB review data were used. In the data sets, two cases, such as the white box and the black box technique, were evaluated.

Keywords: machine learning, white-box, black-box, accuracy.

1 Introduction

The demand for machine learning is now high, which has resulted in a bang in the number of machine learning tools. The number of machine learning libraries and frameworks is so extensive that anyone can apply machine learning techniques to datasets. This is where the problem arises. People who are not experts in this field can blindly apply ML algorithms to their datasets, even sometimes it is done by experts also. The machine learning platform cannot be secure, even if it is secure, the algorithm provider cannot be trusted. As a result, information about the training datasets can be leaked, which can be a huge loss for a company or an organization. The researcher showed that even if the model has a very good accuracy, relatively minor modification can cause information leak. The researchers attacked the four datasets in this paper with two popular cases, such as white-box and black-box. In the case of a white-box attack, the researcher encodes information on the train data in the parameter, since in the white block case, the attacker is aware of the parameter of the trained model. On the other hand, in a black box attack, the attacker has no idea about the parameter, but can access the prediction API. In short, what the researcher wants to show is that using third party code on sensitive data to train models can be dangerous.

2 Background Study

Few terms were used throughout the research report to support the study such as data augmentation, regularization, validation, linear model, deep learning model, etc. The term data augmentation refers to methods for constructing iterative optimization or sampling algorithms via the introduction of unobserved data or latent variables [2] . When training a machine learning model, data augmentation helps minimize overfitting. In this paper, prior to training the data, researchers used data augmentation technique as a preprocessing step to improve the generability of the ML model. Regularization helps to minimize the machine learning model's complexity [3]. Let's say, we want to predict the age of abalone and for that we have a dataset. We now train a model and the model predicts results with 0.995 accuracy, but we only get 0.60 accuracy when we try to run the model on unseen abalone data. That means it results in high variance when we try to train the original dataset because it fits so well with the data pattern that also includes noise. For this reason, from train data to unseen data, the model generalizes poorly. This is referred to as overfitting. To solve the overfitting problem, we need a regularization technique that learns the data pattern and ignores the noise in the data. To do this, the loss function is penalized in order to solve the overfitting problem. Validation is the technique of evaluating a trained model with test data. In this paper, the researcher defines 'accuracy' of a model, f_θ relative to some dataset D using 0-1 loss [1].

$$\mathrm{acc}(\theta, D) = \sum_{(x,y)\in D} \frac{\mathbb{I}(f_\theta(x) = y)}{|D|}$$

Here the function is I outputs 1 if f_θ(x) = y and otherwise output 0 [1]. By measuring the test accuracy, a train model can be validated. The difference between train accuracy and testing accuracy, known as the train-test gap, helps measure how overequipped the model is. Linear models such as support vector machine and logistic regression and deep learning models such as convolutional neural network and residual network were applied in the 5 datasets mentioned in the abstract.

There are many cloud services like the Google Prediction API, Amazon ML, and Microsoft Azure ML that are available in the market. These are known as the machine learning platform. These are the most secure cloud services. The problem occurs when users pay for access algorithms developed by third party clients and there is no guarantee that these third party clients can be trusted even though the platform is the most secure.

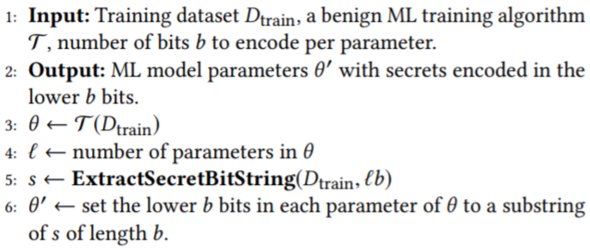
3 Approaches

3.1 Threat Model

Suppose the client now has data and is trying to predict an outcome from the dataset. So he used a training algorithm that was provided by a third party. Therefore, in this paper, the researchers are trying to find out if an adversarial ML provider can filter out sensitive training data even if their code is running on a secure platform. The client is the individual who has the dataset but wants to train their dataset with a machine learning model. The adversary is someone or something that controls the training algorithm. The main objective of the adversary is to infer with the dataset of the client. This can lead to a loss of sensitive information about the data set. What if the adversary is unable to observe the client dataset or the resulting model? Then all he can do is force the model to memorize information to ensure that the model passed the validation. There are two access options: white box and black box. In the white box case, the attacker has information about the model, or I can say that he has direct access to the model. In the black box he has no information about the model. He knows input and output like a customer or a client.

3.2 White-Box

Three types of white-box attacks are described in this paper. Figure 1 is the LSB Encoding attack algorithm where the input is the training dataset and the output is the machine learning model parameters. Line 3 describes that the train starts the model using a conventional training algorithm. Then the number of parameters comes in and the bit string is extracted and finally the parameter is post-processed by setting the lower bits in each parameter that produces this modified parameter. Let's say when decoding you just read the lower bit of the parameters and interpret them as secret bits [1]. In Correlating Value Encoding, The attacker wants to encrypt the secret data or information from the model and adds malicious terms to the loss function [1]. This action is performed while the model parameter is being trained. It's easy to decode the sensitive data from a model when all the features are numeric. A brute force search must be performed to decode the text data. Another attack is a Sign Encoding attack. In order to encode the information, its signs are interpreted as a bit sequence, e.g., a positive parameter represents 1 and a negative parameter represents 0. By reading the signs of model parameters, it is possible to restore the secret data from the model.



```
1: **Input:** Training dataset $D_{\text{train}}$, a benign ML training algorithm
   $\mathcal{T}$, number of bits $b$ to encode per parameter.
2: **Output:** ML model parameters $\theta'$ with secrets encoded in the
   lower $b$ bits.
3: $\theta \leftarrow \mathcal{T}(D_{\text{train}})$
4: $\ell \leftarrow$ number of parameters in $\theta$
5: $s \leftarrow$ **ExtractSecretBitString**$(D_{\text{train}}, \ell b)$
6: $\theta' \leftarrow$ set the lower $b$ bits in each parameter of $\theta$ to a substring
   of $s$ of length $b$.
```

Figure 1. LSB Encoding Attack Algorithm [1].

3.3 Black-Box

Attacking a model without proper knowledge is so much more difficult than a traditional white-box attack procedure. Because the attacker cannot know the model parameter. They discussed two black-box attacks. The first one is Capacity Abuse Attack where input is the training dataset and output is the machine learning model parameters that memorize malicious synthetic inputs and their labels. When a model is trained on the dataset, it becomes overfitted to synthetic input. When the attacker applies synthetic inputs to the trained model, leakage information is possible. Another black-box attack they discussed in this paper is known as Synthesizing Malicious Augmented Data. Data such as synthesizing images and synthesizing text data are required for this attack. If the model is overfitted with these inputs, information leakage can happen.

4 Assessments and Results

4.1 Datasets

As mentioned earlier, 5 publicly available datasets were used in the experiment, such as CIFAR10, LFW, FaceScrub, 20 Newsgroups and IMDB movie review datasets. CIFAR10 and LFW contain image data and Facescrub contains the URLs for 100,000 images. 20 newsgroups consist of 20,000 documents and are referred to as 20 newsgroups because documents are classified into 20 categories. IMDB Movie Review is full of sentiment data with binary classes, positive and negative.

4.2 Evaluation Metrics

The following metrics are used to evaluate the results.

- Accuracy.
- Mean Absolute Pixel Error (for decoding images).
- Precision (for decoding texts).
- Recall (for decoding texts).

4.3 White-Box

The maximum number of lower bits b is used for the LSB encoding attack and the results show that for b $<$ 20, the test accuracy is not reduced for most of the datasets [1]. Support Vector Machine was applied to text datasets such as newsgroup and

IMDB movie review dataset. In SVM, when b ≥ 22, the accuracy does not decrease significantly, even the difference in the baseline test accuracy for the News dataset is positive. We can see a significant drop in baseline test accuracy when the deep learning model is applied to the image dataset. The λ_c coefficient for the correlation term is used for Correlated Encoding Value Attack. Whenever λ_c is 0.1, for image data, we can see the negative difference in baseline test accuracy. For the FaceScrub (F) dataset consisting of image URLs, we can see the opposite result, which is a positive difference in the baseline test accuracy. On the other hand, we see the highest test accuracy drop for IMDB records of 0.66% and a smaller 0.16% drop in test accuracy for text datasets. The sign coding attack does not work well in the binary models for image datasets such as CIFAR10 and LFW. We can observe the increase in test accuracy for λ_s here. For the linear model, such an SVM and logistic regression leads to a decrease in test accuracy for both text data sets. There is no increase in test accuracy for the linear model.

4.4 Black-Box

While the paper discussed two black box attacks, it only showed experiments for capacity abuse attack. The capacity abuse attack performed very well in the datasets. We can see the significant difference between capacity abuse attack and other 3 white box attacks. Deep learning models like RES affect the dataset so we can see the highest number of drops in the baseline test accuracy. For example, we can see the test accuracy decrease for the FaceScrub dataset by 3.72%. CNN shows the increasing baseline test accuracy for LFW dataset. On the other hand, there is no positive baseline test accuracy difference for the linear model when a capacity abuse attack is applied, but a linear model like SVM shows better results than logistic regression. Surprisingly, SVM for the News dataset provides 100% results in both precision and recall. The capacity abuse attack shows that the deep learning model is very good at learning the malicious task.

5 Similar Works

Some researchers from Italy show that it is possible to derive unexpected but useful information from machine learning classifiers [4]. To do this, they created a meta-classifier and trained it to hack other classifiers to get their confidential information from their training datasets. They attacked the classifiers implemented via SVM and the Hidden Mark Model. This way they could determine the leakage of information. Another research paper shows how to exploit sensitive values along with the prediction and to do that they implemented a new class of model inversion attack [5]. The researcher examines decision trees and neural networks. So much research has been done on the subject, and in this modern world, losing information from training data isn't that difficult.

6 Conclusion

This paper discussed the different types of attacks that can harm the machine learning model. The researchers of this paper show how the attacker can lose confidential information from the dataset by slightly changing the training algorithm. The main aim of this research is to show that the machine learning model cannot be used blindly when third party vendors are connected as it can lead to information leaks. The researcher of this paper suggested applying 'the principle of least privilege' to machine learning. So it's easy to say that even if the machine learning platform is secure, don't trust the third party client that provided the algorithm.

Acknowledgements

I would like to thank Dr. Yang Zhang and his team for offering this seminar. It helps me expand my knowledge of data privacy in machine learning. I think the knowledge I have gathered from this seminar will help me in my studies and work in the future.

References

[1] Song, C., Ristenpart, T. and Shmatikov, V., 2017, October. Machine learning models that remember too much. In Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security (pp. 587-601).

[2] Van Dyk, D.A. and Meng, X.L., 2001. The art of data augmentation. Journal of Computational and Graphical Statistics, 10(1), pp.1-50.

[3] Schölkopf, B., Smola, A.J. and Bach, F., 2002. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

[4] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. IJSN, 10(3):137–150, 2015.

[5] Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015.