CIS 6930: Trustworthy Machine Learning

Final Project Report: Title

Manas Gupta
(Point of Contact)
manasgupta@ufl.edu

Malvika Ranjitsinh Jadhav
jadhav.m@ufl.edu

Swarnabha Roy
roys@ufl.edu

December 10, 2021

1 Introduction

With advancements in computational resources, Machine Learning(ML) has encroached into various domains like image classification, speech recognition, generating realistic images and so on. Because of this efficiency and robustness of ML models, ML as a service (MLaaS)[5] is now being offered as part of cloud computing services by companies like Google, Amazon etc. One main reason behind the success of ML models is the availability of large datasets from which these models are trained. Since the data used to train ML models can come from various private sources, vulnerability of these models towards privacy attacks should be minimum.

However, studies have shown that various attacks on ML models can leak sensitive information. One such attack is the Membership Inference Attack or MIA[7]. Given a data sample, a MIA will aim to identify if the target sample is a part of the data used to train the ML model. Thus, removal of sensitive leakable information with minimal compromise on efficiency of the model is important.

During this semester long project, we aimed to compute the loss of membership leakage of data samples i.e., figuring out if a data sample if a part of the training data or not. More specifically, this membership leakage information was calculated for the data points which were part of the model's data at some point but were later removed. This removal of data, also known as machine unlearning, was carried out using the SISA framework. Machine unlearning is important in many practical scenarios as the data used to train the machine learning models comes from multiple sources. So if one or more of the data contributors want to revoke the model's right to use their data, that part of data should be removed from the entire database. Not only this, but the information learnt by the model from this data should also be *unlearnt*.

2 Background & Related Work

2.1 Membership Inference Attacks

In this section we will briefly discuss the membership inference attack we have used in our implementation.

Salem et al. [6] proposed a membership inference attack that relies on using the base model's predictions on the target points. In this scenario, the base model itself serves as its shadow model to approximate the behavior of the model under attack.

In the same paper, the authors also proposed the posterior membership inference attack proposed by Salem et al. [6], which relies mainly on the outcomes of the target model (posterior). Given a target record and a threshold, we query the record on the target model to obtain its predicted probability over the true class label and then decide if it is part of the training data based on whether the posterior is

greater than the threshold.
Researchers have suggested different methods to deal with one of the key factors of membership information leakage, overfitting. These methods include techniques like data augmentation. Kaya et al. [3] in their study investigated the various scenarios and provided guidelines as to when data augmentation can be helpful in reducing MIA leakage.

2.2 Training with Differential Privacy

Differential Privacy (DP) is a framework used to measure the privacy guarantee of an algorithm. Training machine learning models with differential privacy minimizes the risk of exposing sensitive training data. One such approach is the Differentially Private Stochastic Gradient Descent (DP-SGD) [1] where we modify the gradients used in Stochastic Gradient Descent (SGD) based training.

Parameters for DP-SGD:
1. learning rate - This is the learning rate of the SGD training algorithm.
2. number of micro-batches - The CIFAR-10 dataset has 5 batches. However, using the batches is computationally expensive. So we need to break them down into smaller even-sized micro-batches. More number of micro-batches increases the utility but slows down the training process as well.
3. L2 norm clip - This hyperparameter bounds the optimizer's sensitivity to individual training points by clipping the cumulative gradient across all network parameters from each micro-batch to a maximum L2 norm value.
4. Noise Multiplier - To smoothen the dataset we add some amount of noise. Adding more noise ensures better privacy, but compromises on model performance. Setting the noise multiplier between 0.4 and 0.7 was found to be a good trade-off.

2.3 Machine Unlearning

In response to rules like Right To Be Forgotten and the overhead of training from scratch, researchers started looking for methods to make the model forget a portion of the training data (Machine Unlearning). An efficient technique for this was introduced by Lucas Bourtoule [2]. They proposed the SISA (Sharded, Isolated, Sliced, and Aggregated) training approach. This technique focuses on reducing the influence of individual data points on the trained model. The training data is divided into multiple disjoint shards such that only one shard contains a particular training point. Next, we train models in isolation on each of these shards. The data in each shard is divided into slices and presented incrementally during training. Finally, we retrain only the affected model to unlearn a training point.

2.4 Privacy Analysis

The Rényi Differential Privacy (RDP) [4] framework can be used to perform privacy analysis, and measure the differential-privacy guarantee achieved by the model.

Delta and Epsilon are two parameters to show the differential privacy guarantee of our model.
Delta (δ) - It is a bound on the probability that the privacy guarantee of our model does not hold. Generally, this value is set to be less than the inverse of the training data size. We have set it to 10^{-5} because CIFAR-10 has 50000 training points.
Epsilon (ϵ) - This can be called the privacy budget. It bounds how much the probability of the output of our model can vary if we include or remove a particular training example. An epsilon value less than 10 is desirable, although a larger value can still mean good privacy.

3 Approach: Dataset(s) & Technique(s)

3.1 Dataset: CIFAR-10

For our experiments we have used the CIFAR-10 dataset [1]. The dataset consists of 60000 32x32 colour images. These images belong to 10 classes that are mutually exclusive. There are 6000 images per class. The training dataset consists of 50000 images and the testing dataset of CIFAR-10 has 10000 images.

3.2 Implementation technique

Our experiments aim to test whether unlearning techniques can be used to preserve membership privacy. We have implemented our project in three main steps described as follows:

3.2.1 Step 1: Membership Inference attack on Base Model

For our experiments we are using a Resnet based Convolutional neural network mentioned in figure 1

After implementing the above architecture for CIFAR-10 training dataset we implemented the posterior membership inference attack on this model and recorded the results.(Ref: figure 2)

3.2.2 Step 2: Machine unlearning using SISA training

In this step we perform unlearning of data samples using the Sharded, Isolated, Sliced, Aggregated training approach(figure 3).We have divided the training dataset into 10 subsets that we call shards. Each shard consists of 5000 samples. A replica of base model is trained for each shard seperately. These replicas are called constituent models. The 10 shards are further divided into 20 slices each on which the constituent models are incrementally trained incorporating one more slice in the next iteration each time. The predictions are obtained by aggregating results of all the constituent models similar to an ensemble model.

Our program randomly chooses slices in different shards from which data is unlearnt. Thus at the end of this stage we obtain a new model that has unlearnt samples from its training data.

3.2.3 Step 3: Membership inference attack on model after unlearning

After we obtain a model that has unlearnt samples we perform membership inference attacks for the samples that were unlearnt to determine whether machine unlearning plays a role in preserving membership privacy for machine learning models(figure 4). We learnt the posterior attack on the models trained using the SISA framework. Usually, the aggregation of the output of models in the SISA framework is taken as the most common output given by all the constituent models. But in this case since posterior attack requires not just the true label but also its confidence values, we take mean of the posteriors given out by the constituent models.

4 Results

In order to test our approach, we first trained our base architecture(architecture defined in figure 1) on the training samples of the cifar-10 dataset. After training this model for 20 epochs, we got a training accuracy and loss of 99.99% and 0.0007 respectively while accuracy and loss on a handout dataset were 89.7% and 0.57 respectively. Next, we ran posterior attack on our architecture by sampling random 2000 samples from our training and testing datasets. After varying the value of threshold t, the best attack accuracy and advantage on the base architecture were found out to be 85% and 0.17 respectively.
Machine unlearning was carried out using the SISA framework described previously. Constituent models

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

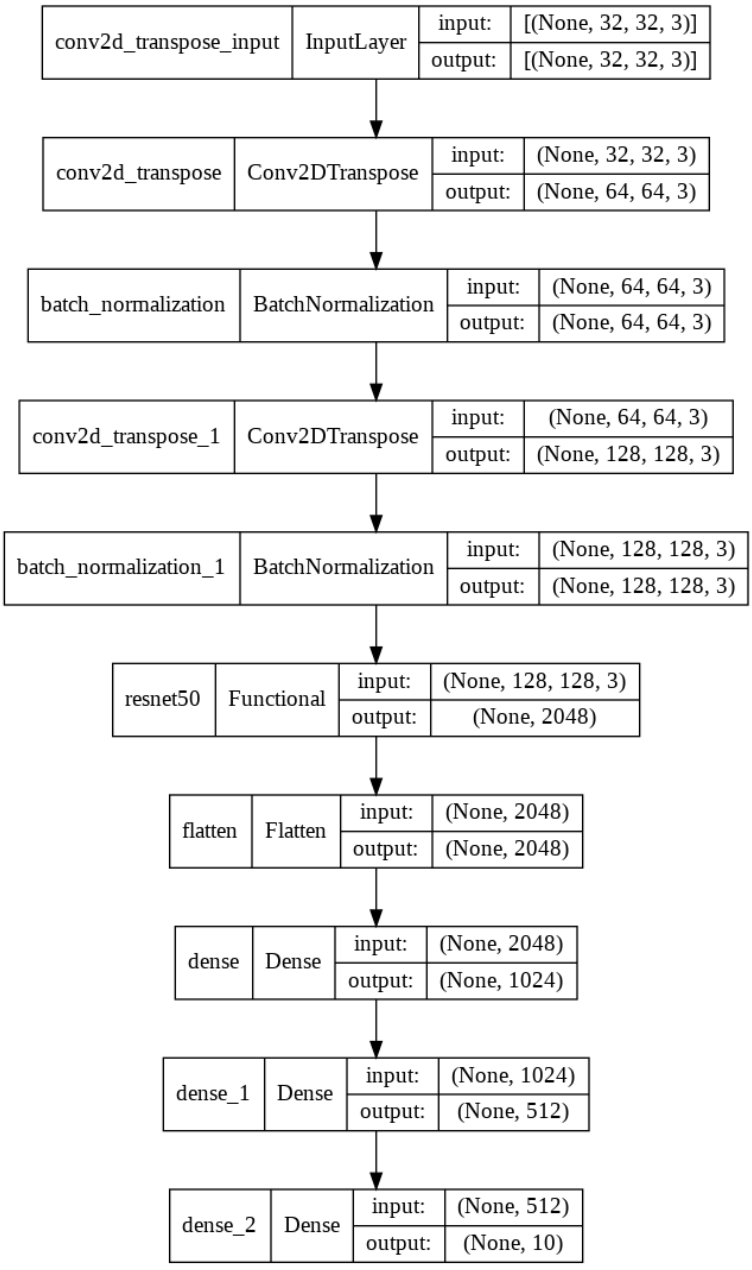

Figure 1: Architecture of Base model

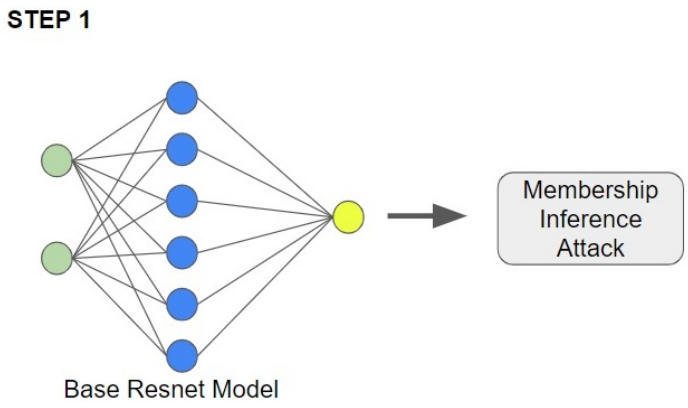

Figure 2: Step 1

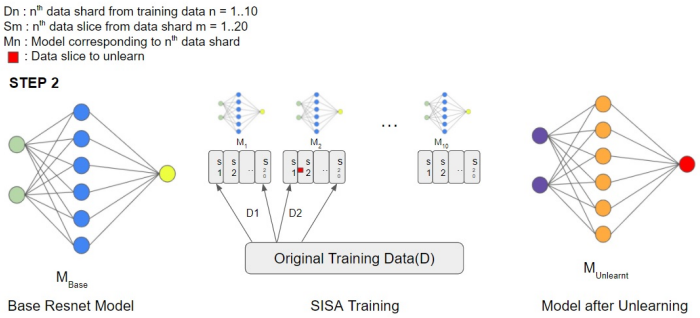

Figure 3: Step 2

created during this process gave an training accuracy of 82% and testing accuracy of 68%. These unlearnt models were later subjected to posterior attack which resulted in an accuracy of 52.6% and advantage of 0.05. More details about the implementation of various parts of this project have been mentioned in the Appendix.

5 Conclusions

Previous works have focused on different techniques like differential privacy for obtaining privacy guarantee. Through the project we aimed to do a novel investigation to find out whether machine unlearning plays a role in preserving membership privacy. During our implementation we also modified the aggregation step in SISA training to generate output in terms of confidence values opposed to a target label. Our experimental results show that unlearning techniques can bring the membership inference attack accuracy very close to random guessing. Our results also point out that using a training method that works with small sized data shards can result in compromising accuracy of overall predictions especially if performed on overparameterized models. Thus machine unlearning can be used to preserve membership privacy for machine learning models.
It should also be noted that in this work, we have focused our experiments on membership privacy of resnet based CNNs. Generalization of these results needs verification across a broad class of networks which was out of the scope for this project.

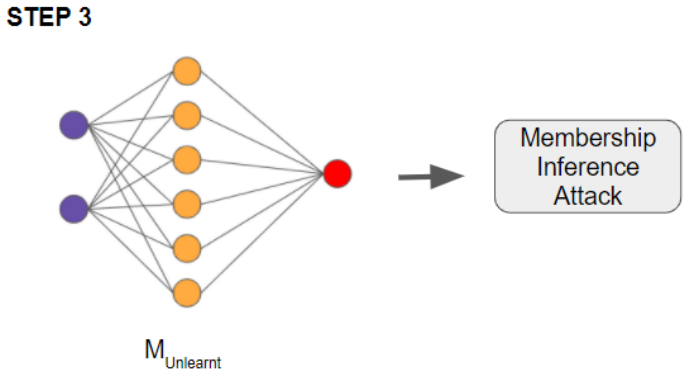

Figure 4: Step 3

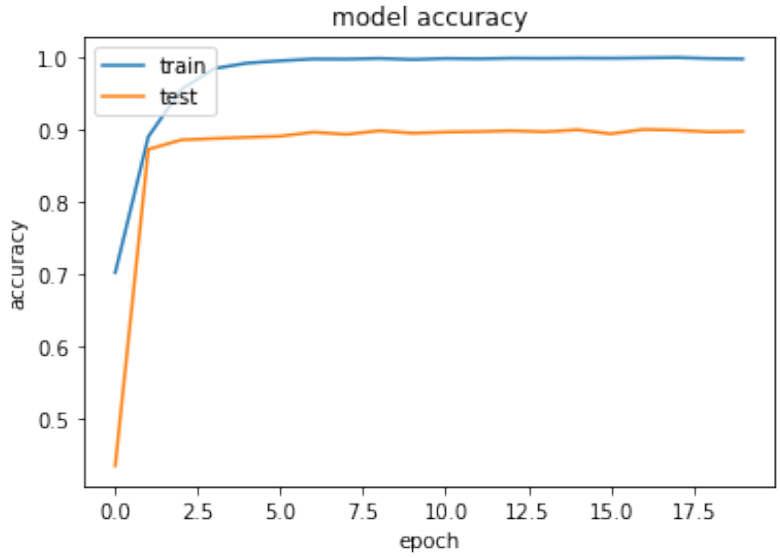

Figure 5: Running time vs length of string generated

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.

[3] Yigitcan Kaya and Tudor Dumitras. When does data augmentation help with membership inference attacks? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5345–5355. PMLR, 18–24 Jul 2021.

[4] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

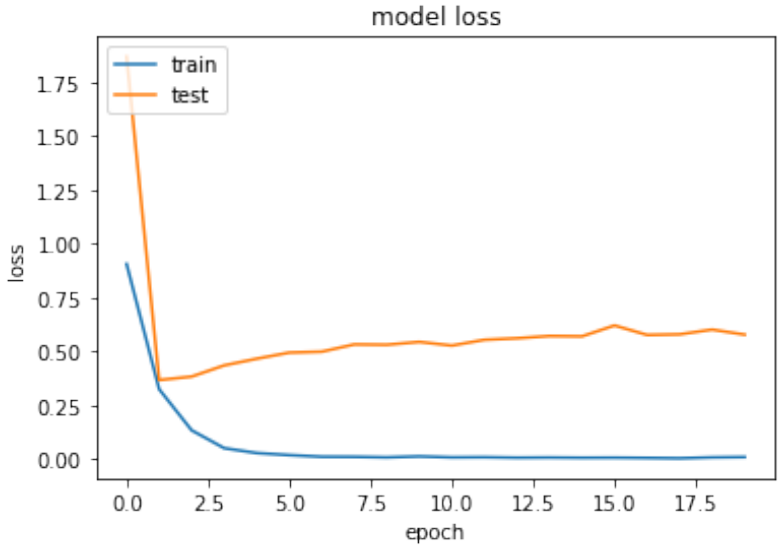

Figure 6: Running time vs length of string generated

[5] Mauro Ribeiro, Katarina Grolinger, and Miriam A. M. Capretz. Mlaas: Machine learning as a service. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 896–902, 2015.

[6] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

[7] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

6 Appendix

6.1 Interesting takeaways during the implementation:

[1]We have implemented DP-SGD training and evaluated it using the RDP framework. However, this gives us an idea of the overall privacy guarantee of the entire dataset, and not a target data point. Thus, we could not use it as a metric to know if our unlearnt model has actually unleanrt the intended data.

[2]We had tried implementing the Logit based linear filtration method of unlearning but we discovered the method focus on filtering out the outputs of data points for which unlearning is to be done. While this approach can ensure that membership privacy of the data point is preserved it does not perform unlearning of the data point from the trained base model.