Experiments of the membership inference attack

Shin, Jongho

April 2019

1 Introduction

This report presents experiments on the membership inference attack of Shokri et al.[1] The experiments examined the viability of the attack and some dominant factors of this attack. For the experiments, 2 datasets have been chosen from the paper: CIFAR-10, and UCI Adult income data. The experiments had conducted using Keras and Tensorflow for the neural networks, and scikit-learn for SVM models.

The code is available in Github: https://github.com/Jongho0/ml_mbr_inf

2 Experiment

2.1 Dataset

CIFAR-10. CIFAR-10 is the main dataset of this report. CIFAR-10 has 10 classes with 6,000 image samples per class. Each sample is 32x32x3 image data.

UCI Adult (Census Income). This dataset is from the 1994 Census database, and composed of 14 attributes and 2 classes; $> 50K$ or $<= 50K$. 3,000 of randomly chosen records were used to train the target model.

2.2 Experiment Environment

To reproduce the same result for CIFAR-10, I mimicked the original settings. Thus I varied the training size to 2,500; 5,000; 10,000; and 15,000 for each number of classes. For this reproduction, the number of shadow models was 20, and the number of epochs was 30. The target model was a CNN with 2 convolution and max-pooling layers and 2 hidden layers. To examine the effect of the number of shadow models, I varied the number of shadow models to 1, 10, 25, 50, 70, and 100, while training size has set to 3,000. I also varied the number of epochs to see the effect of overfitting. The number of epochs was set to 10, 50, 100, 200, 300, and 500. For UCI Adult dataset, the number of epochs was 100, and the training size was 5,000. The target model was an FCNN with 4 hidden layers.

For the attack model, I used SVMs for each class.

2.3 Accuracy of the attack

Dataset	Training Accuracy	Testing Accuracy	Attack Precision
CIFAR-10	1.0000	0.432	0.8284
Adult	0.9747	0.785	0.522

Table 1: Accuracy of the target models and the corresponding attack precision

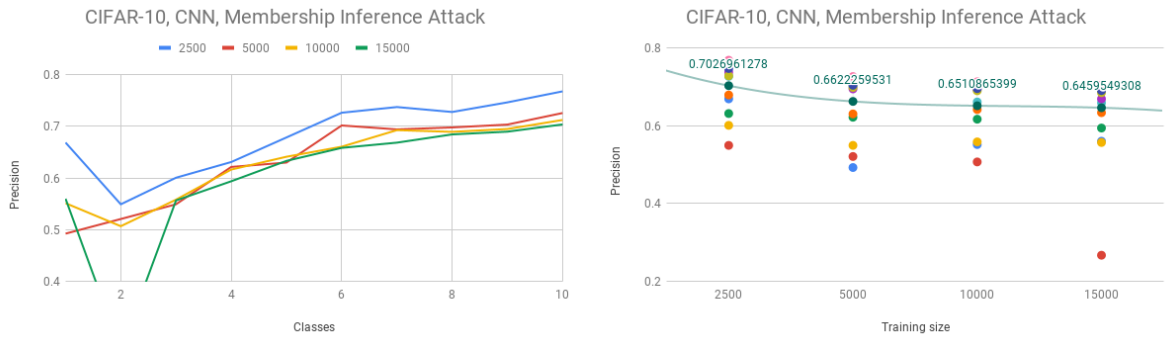

Figure 1: Precision of the membership inference attack against neural networks trained on CIFAR datasets.

Table 1 shows the accuracy of the target models and the attack precision against them. In the case of CIFAR-10, the attack was very successful with high precision. For UCI Adult income data, the attack was slightly better than the baseline. Comparing to table II of the original paper, however, this experiment shows better precision for the Adult dataset; it was 0.503 in the original paper. I think it is due to the bigger accuracy gap between the training and the testing.

2.4 Effect of the number of classes and training data per class

For the reproduction of the original experiment, figure 1 shows the result of the membership inference attack against the CIFAR models, and it is similar to figure 4 of the original paper. Even though the attack worked poor in the case of 2 classes, the attack performed well above the baseline. As the original paper, the graphs show that the precision tends to decrease as the training size increases. Moreover, the first graph shows a better correlation between the number of classes and precision than one in the original paper.

2.5 Effect of the number of shadows

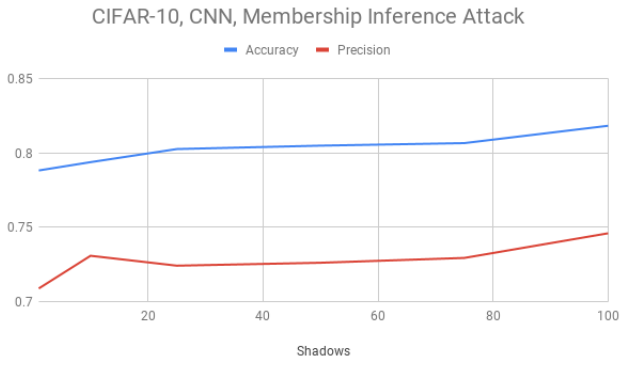

Figure 2: Accuracy and precision with a different number of shadow models

Figure 2 presents the precision and recall of the various number of shadow models. It is not dramatic, but this graph shows a definite positive correlation between them. It seems obvious because more shadow models mean more training data for the attack model. I think the slope is relatively flat due to the training set size of the attack model. I kept the training set size of the attack model to 2,000 for all the different

number of shadow models. Since the training set size is relatively small, it might be a bit difficult to embrace the diversity added by new shadow models.

2.6 Effect of overfitting

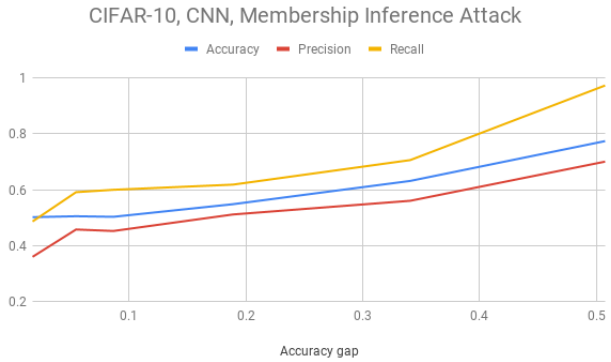

Figure 3: Accuracy, precision, and recall for different accuracy gaps between training and testing

Figure 3 shows the effect of overfitting. I varied the number of epochs and measured the accuracy gap between the training set and test set as well as the precision of the attack. For this experiment, I used the stochastic gradient descent as the optimizer for slow learning. This graph presents the clear relationship of overfitting and the attack precision. So their assumption, 'The more overfitted a model, the more it leaks', seems plausible.

3 Conclusion

From the experiments, I was able to reproduce the result of the original paper and confirm the validity of the membership inference attack. In addition, the effects of shadow models and overfitting have been presented more clearly. Since the membership inference attack is valid for many neural networks, in-depth studies of its characteristics and following defense strategies will be required.

References

[1] 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.