Brandon Edmunds
Pruning/Privacy
Dr. Mahdavi and Dr. Maeng
12/02/22

Privacy Evaluation Summary

Secret Sharer ([1])

Secret Sharer proposes measuring privacy as a model's ability for unintended memorization. Unintended memorization is characterized by the extent to which a model memorizes data that is out-of-distribution. The intuition is that a model is likely interested in memorizing data that fits the target distribution trying to be modeled, and that the model should not be interested in memorizing data that is not helpful in learning the target distribution, with such data being considered as out-of-distribution, also called canaries. To evaluate a model, a model is trained with additional canaries added to the initial data, then the exposure of the canary is measured. The exposure is a metric for generative models that measures the extent of information of a canary available given the model. The importance in the model being generative is because the exposure metric is defined assuming the canary is a part of the data sample space so that the model can generate the canary as the output.

The paper also empirically determines that random dropout and quantization do not affect unintended memorization, while DP-SGD decreases unintended memorization.

DP-Auditing ([2])

[2] investigates the practical implications of differential privacy in determining the information necessary to the attacker to achieve bounds on ε and δ. Figure 1 illustrates the DP-auditing done in [2].

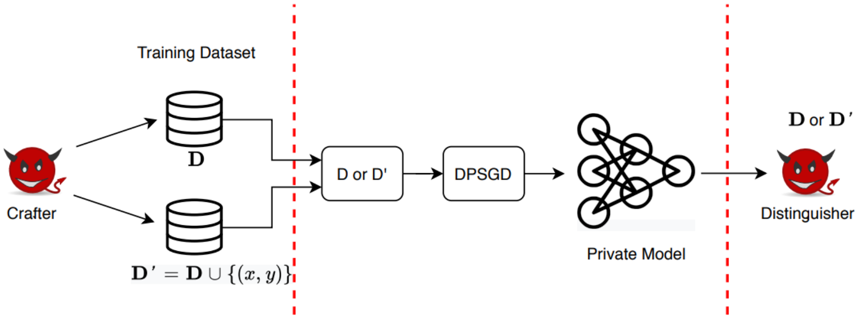

Figure 1 – Adversary Instantiation ([2])

The paper focuses on empirically determining the differential privacy given different parameters in Figure 1, such as having crafters and distinguishers with access to different amounts of information or using different values of ε in DPSGD. The sequence of inserting an additional element into the original dataset, training on the dataset, and having the distinguisher guess which dataset was used is done many times (1000 in the paper), then the false positive and false negative rates are used to determine the resulting ε and δ.

References

[1] https://www.usenix.org/system/files/sec19-carlini.pdf
[2] https://arxiv.org/pdf/2101.04535.pdf