A flaw in the NeurIPS unlearning challenge and an algorithmic framework for entropy regularization.

Sumeet Shirgure
University of Southern California
Los Angeles, California
sshirgur@usc.edu

Abstract

This paper documents some findings from experiments pertaining to the NeurIPS unlearning challenge. [1] The contributions of this paper are as follows : the importance of entropy of the outputs of deep neural network classifiers is discussed as a means of conducting membership inference attacks Hu et al. (2022), and a strategy to evade such entropy based attacks is provided in what's called the "obfuscation" framework. The experiments evaluate the performance of a model constructed using the proposed framework vs. the oracle model given in the starter kit of the NeurIPS challenge against such attacks. We subsequently show that considering the oracle model's output distribution as currently defined is likely not the best approach to define the forgetting quality of an unlearning algorithm. The proposed algorithmic framework can be of independent interest.

1 Index

The reader is referred to the link in the footnote to first learn about the setting and the setup of the unlearning challenge; we assume that the reader is familiar with terms like "retain" and "forget" sets. The rest of the paper is structured as follows : section 2 rehashes existing membership inference attacks that consider the uncertainty of the neural network in predicting a class label; section 3 will look at an algorithmic framework to mitigate against membership inference attacks; section 4 goes over experiments, results, and concludes that the oracle model as defined has its flaws.

2 Entropy of classifier outputs

Deep neural networks tend to have lower values of the cross-entropy loss, which they were trained to minimize, for inputs from the training set Hu et al. (2022), Shokri et al. (2017) when compared to inputs not in the training set. In the unlearning challenge, this trend translates to the average loss of the oracle model being lower for samples from the retain set compared to that of the unseen and forget data subsets. What we will show is that the ideal model is also overfit in some sense to the retain dataset. An adversary can simply train a logistic regression model on the output loss to differentiate between a data point from the retain set vs. one not in it.

All of the above is true for the average entropy of the output as well (and not just cross entropy with the target distribution). That's precisely what an entropy based attack does : if the pure entropy of the output logits of the neural network is below a certain threshold, it classifies the point as from the training dataset. The advantage of studying entropy attacks and not just cross entropy attacks is that the *latter depends on the class labels*, whereas the former has no such restriction and purely depends on the input and the deep neural network model weights. Salem et al. (2018).

[1] https://unlearning-challenge.github.io/

3 An algorithmic framework for obfuscation of classifier outputs

The contents of the last section hint at the need for a regularization technique that mitigates against overfitting just on the retain dataset. We will now look at one such strategy. The idea is to penalize the KL-divergence between the output distribution and the uniform distribution.

$$D_{KL}(P\|\mathcal{U}) = \sum_i p_i(log(p_i) - log(\frac{1}{N})) = -H(P) + log(N)$$

Where N is the number of classes. Note that this is the same as a regularizing term that penalizes small output entropy. We note that such methods have already been studied, for example in Shokri et al. (2017), where the broadening of probability distribution over class labels is achieved by increasing the temperature of the softmax layer on the model's outputs as introduced in Hinton et al. (2015). And using KL-divergence as a regularization term is also not a new concept. E.g it's widely studied in variational autoencoding. Kingma and Welling (2022).

We now define an (ϵ, δ)-obfuscation as a procedure with two phases : (a) first, perform gradient updates corresponding to D_{KL} with some rate η : $\theta' \leftarrow \theta + \eta\nabla_\theta(D_{KL}(P\|\mathcal{U}))$ until the standard deviation of the entropy of the output $H(P)$ over the training set drops below ϵ. (b) then optimize for $\mathcal{L} + \alpha D_{KL}(P\|\mathcal{U})$ until validation accuracy is $1 - \delta$ times that before phase (a), where $\mathcal{L}$ is the pre-obfuscation loss function.

η, α, ϵ and δ are tunable hyperparameters. Note that the accuracy of the classifier can drop arbitrarily between phases and (a) and (b), and there's no guarantee that either phase terminates for any given ϵ, δ. But empirical evidence suggests that for reasonable values of these hyperparameters, phases (a) and (b) should terminate quickly and phase (b) requires far fewer epochs than training from scratch. Essentially this framework allows us to "convert" models that were not originally trained with regularization into those that have broad distributions without "retraining a new model from scratch" so to speak. We ask the community for a theoretical study of this framework.

4 Experimental setup and results

For the unlearning challenge, the results of the experiments are given in table A. The original model is a ResNet18 pretrained on CIFAR10 (retain+forget sets), the oracle model is trained solely on the retain subset, and the obfuscated model is the original model after undergoing obfuscation. As can be seen in the table, the obfuscated model evades both kinds of membership inference attacks, one between retain and forget sets, and one between forget and test(unseen) sets. And it is just as good at classifying images from the test set. Hopefully this is enough to throw shade on the choice of the oracle model in the unlearning challenge. E.g in the unlearning challenge, had the organizers declared that the oracle model will be trained with regularization the reference distribution would have changed drastically while still performing at comparable accuracy. To see this in practice, please refer to the histograms generated in the attached Jupyter notebooks in the appendix. We further conduct similar experiments on various classifiers trained on datasets like CIFAR10, CIFAR100, and other choices of model architectures like the vision transformer Dosovitskiy et al. (2021) that show the utility of the proposed framework.

Summary

We have explored the role of the entropy of the output distributions of deep neural network classifiers in their susceptibility to entropy based membership inference attacks. We then proposed a framework for obfuscating said distribution by regularization of entropy to the point that it concentrates around an ϵ sized region, followed by retraining with regularization. We then applied the obfuscation algorithm to pretrained models in the NeurIPS unlearning challenge to get a model that's better than the oracle model in the sense that it evades membership inference attacks between retain and forget sets. We also ran similar experiments after changing network architectures and datasets to observe consistent trends.

References

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s), sep 2022. ISSN 0360-0300. doi: 10.1145/3523273. URL `https://doi.org/10.1145/3523273`.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

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, 2018.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models, 2017.

A Appendix

Metric	Original model	Oracle model	**Obfuscated model**
Test set accuracy	88.3%	88.0%	88.2%
Average retain set entropy	0.047 ± 0.124	0.043 ± 0.116	2.293 ± **0.004**
Average forget set entropy	0.048 ± 0.123	0.145 ± 0.282	2.293 ± **0.004**
Average test set entropy	0.137 ± 0.278	0.143 ± 0.280	2.293 ± **0.004**
Retain vs. forget MIA acc.	N/A	58.6%	**50.2%**
Forget vs. unseen MIA acc.	57.6%	49.7%	**51.7%**

Table A : results from unlearning challenge

Note that the obfuscated model is the only one whose entropy doesn't vary much. This makes it resistant to entropy or cross entropy based membership inference attacks.

The reader is referred to the following link containing notebooks for more details.

[Anonymous Google drive link]

`https://drive.google.com/drive/folders/1Kp4wyvjDyAJG4-SdtW68N53YzKs-DMyu?usp=sharing`

In the MNIST notebook, note that simply training without regularization is enough for MNIST classifiers to evade MIAs. This suggests that the framework is only useful for rich datasets where overfitting and dataset membership information leakage are an issue.

In the CIFAR-100 notebook, a similar resnet classifier is trained and then obfuscated on CIFAR100.

In the ViT notebook, a vision transformer is trained on CIFAR10.