TMI! Finetuned Models Spill Secrets from Pretraining

John Abascal, Stanley Wu, Alina Oprea, Jonathan Ullman
Northeastern University

Paper

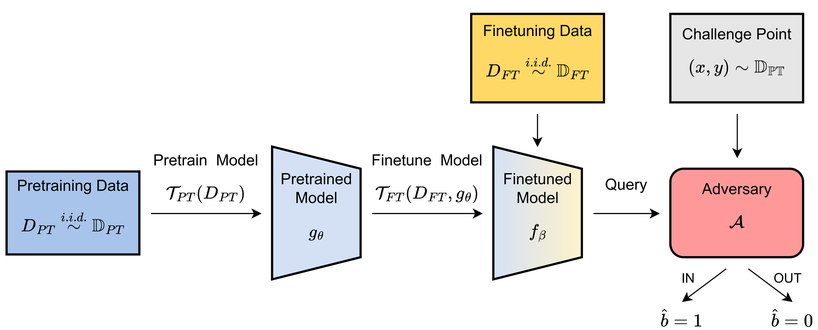
Contact

Introduction

- **Transfer learning** has become an increasingly popular technique in **private machine learning** as a way to leverage a model trained for one task to assist with building a model for a related task.
- Because pretraining data is often considered to be public, it provides a good initialization for sensitive downstream tasks.
  - Unfortunately, large, public datasets are typically scraped from the Web indiscriminately, raising concerns about the sensitivity of this data.

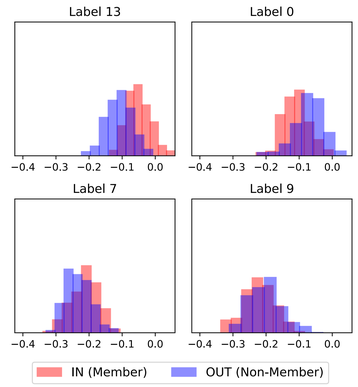
Thus, the central question we attempt to understand in this work is: **How much sensitive information does a *finetuned* model reveal about the data that was used for *pretraining*?**

Membership Inference Threat Model



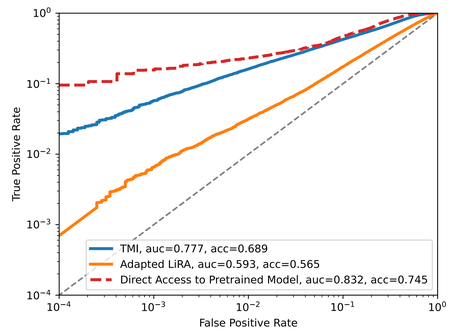
Methodology

- We propose a metaclassifier-based **membership inference attack**, **TMI.**
  - TMI leverages the influence of memorized pretraining samples to infer membership by making a metaclassification over all possible labels in the downstream task.

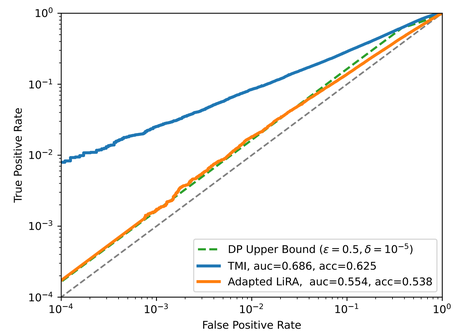


Logit Distributions of Finetuned Models Queried on Pretraining Data

Results



Finetuned on Coarse-Labeled CIFAR-100



Finetuned on Coarse-Labeled CIFAR-100 *with DP* $(\varepsilon = 0.5, \delta = 10^{-5})$ **when** $D_{PT} \cap D_{FT} \neq \emptyset$

Conclusion

- **Finetuning leaves pretraining examples at risk of membership inference attacks.**
- If an individual's data was included in a public pretraining dataset, finetuning on this individual's data with DP will ***not*** protect them from membership inference attacks.

Task	TPR @ 0.1% FPR	TPR @ 1% FPR
TMI (Coarse CIFAR-100)	5.7%	16.1%
TMI (CIFAR-10)	2.0%	8.0%
Adapted LiRA (Coarse CIFAR-100)	0.7%	3.1%
Adapted LiRA (CIFAR-10)	0.3%	1.5%
LiRA Directly on Pretrained Model	15.6%	22.9%

Task	TPR @ 0.1% FPR	TPR @ 1% FPR
TMI - Coarse CIFAR-100; $(\varepsilon = 0.5, \delta = 10^{-5})$	2.6%	8.5%
TMI - Coarse CIFAR-100; $(\varepsilon = 1, \delta = 10^{-5})$	4.2%	12.6%
Adapted LiRA - Coarse CIFAR-100 $(\varepsilon = 0.5, \delta = 10^{-5})$	0.17%	1.8%
Adapted LiRA - Coarse CIFAR-100 $(\varepsilon = 1, \delta = 10^{-5})$	0.35%	2.8%
DP Upper Bound $(\varepsilon = 0.5,\ \delta = 10^{-5})$	0.16%	1.6%
DP Upper Bound $(\varepsilon = 1,\ \delta = 10^{-5})$	0.27%	2.7%

Attack Performance