Membership Privacy for Machine Learning Models Through Knowledge Transfer

Virat Shejwalkar and Amir Houmansadr
University of Massachusetts Amherst
{vshejwalkar, amir}@cs.umass.edu

Abstract

Large capacity machine learning (ML) models are prone to membership inference attacks (MIAs), which aim to infer whether the target sample is a member of the target model’s training dataset. The serious privacy concerns due to the membership inference have motivated multiple defenses against MIAs, e.g., differential privacy and adversarial regularization. Unfortunately, these defenses produce ML models with unacceptably low classification performances.

Our work proposes a new defense, called distillation for membership privacy (DMP), against MIAs that preserves the utility of the resulting models significantly better than prior defenses. DMP leverages knowledge distillation to train ML models with membership privacy. We provide a novel criterion to tune the data used for knowledge transfer in order to amplify the membership privacy of DMP.

Our extensive evaluation shows that DMP provides significantly better tradeoffs between membership privacy and classification accuracy compared to state-of-the-art MIA defenses. For instance, DMP achieves ~100% accuracy improvement over adversarial regularization for DenseNet trained on CIFAR100, for similar membership privacy (measured using MIA risk): when the MIA risk is 53.7%, adversarially regularized DenseNet is 33.6% accurate, while DMP-trained DenseNet is 65.3% accurate. We have released our code at github.com/vrt1shjwlkr/AAAI21-MIA-Defense.

Introduction

The remarkable performance of machine learning (ML) in solving many classification tasks has facilitated its adoption in various domains ranging from recommendation systems to critical health-care management. Many ML-as-a-Service platforms (e.g., Google API, Amazon AWS) enable novice data owners to train ML models and release the models either as a blackbox prediction API or as model parameters that can be accessed in whitebox fashion.

ML models are often trained on data with sensitive user information such as clinical records and personal photos. Hence, ML models trained using sensitive data can leak private information about their data owners. This has been demonstrated through various inference attacks (Fredrikson et al. 2015, Hitaj et al. 2017, Carlini et al. 2018), and most notably the membership inference attack (MIA) (Shokri et al. 2017) which is the focus of our work. An MIA adversary with a blackbox or whitebox access to a target model aims to determine if a given target sample belonged to the private training data of the target model or not. MIAs are able to distinguish the members from non-members by learning the behavior of the target model on member versus non-member inputs. They use different features of the target model for this classification, e.g., model predictions (Shokri et al. 2017), model loss, and gradients of the model parameters for given input (Nasr et al. 2019). MIAs are particularly more effective against deep neural networks (Shokri et al. 2017; Salem et al. 2019), because, with their large capacities, such models can better memorize their training data.

Recent work has investigated several defenses against membership inference attacks. In order to provide the worst case privacy guarantees, Differential Privacy (DP) based defenses add very large amounts of noise to the learning objective or model outputs (Papernot et al. 2017, Chaudhuri et al. 2011). This results in models with unacceptable tradeoffs between privacy and utility (Jayaraman et al. 2019), therefore questioning their use in practice. Sablayrolles et al. (2019) showed that membership privacy is a weaker notion of privacy than DP, which improves with generalization of ML models. Similarly, Nasr et al. (2018) proposed adversarial regularization targeted to defeat MIAs by improving the target model’s generalization. However, as we demonstrate, the adversarial regularization and other state-of-the-art regularizations, including label smoothing (Szegedy et al. 2016) and dropout (Srivastava et al. 2014), fail to provide acceptable membership privacy-utility tradeoffs (simply called ‘tradeoffs’ here onward). Memguard (Jia et al. 2019), a blackbox defense, improves model utility, but it cannot protect the model from whitebox MIAs and even the simple threshold based MIAs (Yeom et al. 2018). In summary, existing defenses against MIAs offer poor tradeoffs between model utility and membership privacy.

To this end, our work proposes a defense against MIAs that significantly improves the tradeoffs compared to prior defenses. That is, for a given degree of membership privacy (i.e., MIA resistance), our defense produces models with significantly higher classification performances compared to prior defenses. Our defense, called Distillation for Membership Privacy (DMP), leverages knowledge distillation (Hinton et al. 2014), which transfers the knowledge...
of large models to smaller models, and is primarily used for model compression. Intuitively, DMP protects membership privacy by thwarting the access of the resulting models to the private training data. The first pre-distillation phase of DMP trains an unprotected model on the private training data without any privacy protection. Next, in distillation phase, DMP selects/generates reference data and transfers the knowledge of the unprotected model into predictions of the reference data. In the final post-distillation phase, DMP trains a protected model on the reference data labeled in the previous phase. Unlike conventional distillation, we use the same architectures for the unprotected and protected models.

Similar to adversarial regularization and PATE, DMP assumes access to a possibly sensitive and “unlabeled” reference data drawn from the same distribution as the “labeled” private training data, and uses such reference data to train its final models; the reference data is not publicly available. This is a highly realistic assumption as typical model generating entities (e.g., banks) possess huge amounts of “unlabeled” data (but limited labeled data due to the expensive labeling process). Furthermore, we show that this assumption can be relaxed by synthesizing reference data using generator networks (Micaelli et al. 2019). While some prior work (Papernot et al. 2017) combined distillation and DP to protect data privacy, our work is the first to study the promise of knowledge distillation as the sole technique to train membership privacy-preserving models. Our key contributions are summarized below:

- We propose a defense against MIAs, called Distribution for Membership Privacy (DMP).
- Given an unprotected model trained on a private training data and a reference sample, we provide a novel result that the lower the entropy of prediction of the model on the reference sample, the lower the sensitive membership information in the prediction. We use this result to select/generate appropriate reference data so as to improve the membership privacy due to DMP.
- We perform an extensive evaluation of DMP to show the state-of-the-art tradeoffs between membership privacy and model accuracy of DMP. For instance, at a fixed high degrees of membership privacy, DMP achieves 30% to 140% higher classification accuracies compared to state-of-the-art defenses across various classification tasks.

**Related Work**

**Membership Inference Attacks** Shokri et al. (2017) introduced membership inference attacks (MIAs). Given a target model trained on a private training data and a target sample, MIA adversary aims to infer whether the target sample is a member of the private training data. Shokri et al. (2017) proposed to train a neural network to distinguish the features of the target model on members and non-members. They assumed a partial access to the private training data. Salem et al. (2019) relaxed this assumptions and showed the transferability of MIAs across datasets. These works relied on the blackbox features of target models, e.g., model predictions to mount MIAs. Nasr et al. (2019) proposed to use whitebox features of target models, e.g., model gradients, along with the blackbox features, to further enhance the MIA accuracy. Above works used generalization gap (i.e., difference in train and test accuracy) of target models to mount strong MIAs. The more recent MIA literature focuses on deriving features that can better distinguish the behavior of target models on members and non-members (Leino et al. 2019; Song et al. 2020).

**Defenses Against Membership Inference Attacks** MIAs exploit differences in behaviors of target models on members and non-members. Regularization techniques, including dropout and label smoothing, reduce the difference in terms of accuracies of the target model on members and non-members, and mitigate MIAs to some extent (Shokri et al. 2017). Nasr et al. (2018) proposed adversarial regularization (AdvReg) tailored to defeat MIAs. AdvReg simultaneously trains the target and attack models in a game theoretic manner, and regularizes the target model using the accuracy of the attack model. The final target models that use above regularization defenses can be deployed in whitebox manner, i.e., similar to DMP, they are whitebox defenses. Hence, we thoroughly compare our DMP defense with all these regularization techniques. However, as shown in (Song and Mittal 2020) and seen from the original work (Nasr et al. 2018), AdvReg is not an effective defense, because it either fails to mitigate MIA or incurs large drops in model utility (classification accuracy). Jia et al. (2019) proposed MemGuard, a blackbox defense that adds noise to the output of the target model such that the noisy output is both accurate and fools the given MIA attack model. However, MemGuard does not defend against the simplest of threshold based attacks (Yeom et al. 2018; Sablayrolles et al. 2019). We omit MemGuard and other blackbox defenses, e.g., top-k predictions (Shokri et al. 2017), from evaluations.

Differential privacy based defenses such as DP-SGD (Abadi et al. 2016) and PATE (Papernot et al. 2017) are whitebox defenses and provide strong theoretical membership privacy guarantees. However, as (Jayaraman and Evans 2019) show—and we confirm in our work—target models trained using DP-SGD and PATE have prohibitively low classification accuracies rendering them unusable.

**Preliminaries**

**Knowledge Distillation** Bucilua et al. (2006) and Ba et al. (2014) proposed knowledge distillation, which uses the outputs of a large teacher model to train a smaller student model, in order to compress large models to smaller models. The outputs used for distillation can vary, e.g., Hinton et al. (2014) use class probabilities generated by the teacher as the outputs, while Romero et al. (2014) use the intermediate activations along with class probabilities of the teacher. It is well established that knowledge distillation produces students with accuracies similar to their teachers (Crowley, Gray, and Storkey 2018; Zagoruyko and Komodakis 2016). This also allows DMP to produce highly accurate target models. Note that, although we use term “distillation”, DMP uses teacher and student models of the same sizes, because DMP is not concerned with the size of the resulting model.

9550
Membership Inference Attacks  Below we give the threat model and MIA methodology that we consider in this work.

**Threat model.** The primary goal of the adversary is to infer the membership of a target sample \((x, y)\) in the private training data \(D_{tr}\) of a target model \(\theta\). Our DMP defense uses private, unlabeled reference data \(X_{ref}\) for knowledge transfer, which itself could be privacy sensitive, hence, we consider a secondary goal to infer membership of a target sample in \(X_{ref}\). Following the previous works, we assume a strong adversary with the knowledge of: target model parameters (the strongest whitebox case), half of the members of \(D_{tr}\) and equal number of non-members. Similarly, to assess the MIA risk to \(X_{ref}\), we assume that the adversary has half of the members of \(X_{ref}\) and the equal number of non-members. Note that, the assumptions on the partial availability of private \(D_{tr}\) and private \(X_{ref}\) facilitates the assessment of defenses under a very strong adversary. The adversary can compute various whitebox and blackbox features of the target model and train an attack model. The adversary cannot poison \(X_{ref}\) as it is not publicly available.

**Methodology.** Consider a target model \(\theta\) and a sample \((x, y)\). MIAs exploit the differences in the behavior of \(\theta\) on members and non-members of the private \(D_{tr}\). Therefore, MIAs train a binary attack model to classify target samples into members and non-members. Such attack models can be neural networks (Shokri et al. 2017; Salem et al. 2019) or simple thresholding functions where threshold is tuned for maximum attack performance (Yeom et al. 2018; Sablayrolles et al. 2019; Song and Mittal 2020). The adversary computes various features of \(\theta\) for given \((x, y)\), e.g., prediction \(\hat{y}(x, y)\), \(\theta\)'s loss on \((x, y)\), and the gradients of the loss. The adversary combines these features to form \(F(x, y, \theta)\).

The attack model \(h\) takes \(F(x, y, \theta)\) as its input and outputs the probability that \((x, y)\) is a member of \(D_{tr}\). Let \(Pr_{D_{tr}}\) and \(Pr_{D_{tr}}\) be the conditional probabilities of the members and non-members of \(D_{tr}\), respectively. Hence, the expected gain of the attack model for the above setting is given by:

\[
G^\theta(h) = \mathbb{E}_{(x, y) \sim Pr_{D_{tr}}} \left[ \log(h(F)) \right] + \mathbb{E}_{(x, y) \sim Pr_{D_{tr}}} \left[ \log(1 - h(F)) \right]
\]

In practice, the adversary knows only a finite set of the members \(D\) and non-members \(D'\) required to train \(h\), hence computes the above gain empirically as:

\[
G^\theta_{D, D'}(h) = \sum_{(x, y) \in D} \frac{\log(h(F))}{|D|} + \sum_{(x, y) \in D'} \frac{\log(1 - h(F))}{|D'|}
\]

Finally, the adversary solves for \(h^*\) that maximizes (2).

**Our Proposed Defense: DMP**

Now, we present our defense Distillation For Membership Privacy (DMP), which is motivated by the poor membership privacy-utility tradeoffs provided by existing MIA defenses (§). First, we give an intuition behind DMP and detail the DMP training. Finally, to achieve the desired tradeoffs, we give a criterion to tune the selection or generation (e.g., using GANs) of reference data used in DMP.

**Notations** \(D_{tr}\) is a private training data. An ML model trained on \(D_{tr}\) without any privacy protections is called unprotected model, denoted by \(\theta_{up}\). An ML model is called protected model, denoted by \(\theta_p\), if it protects \(D_{tr}\) from MIAs. For knowledge transfer, DMP uses an unlabeled and possibly private reference dataset which is disjoint from \(D_{tr}\); as the reference data is unlabeled, we denote it by \(X_{ref}\). We denote the soft label of \(\theta\) on \(x\), i.e., \(\theta(x)\), by \(\theta^x\).

**Main Intuition of DMP** Sablayrolles et al. (2019) show that \(\theta\) trained on a sample \(z\) (short for \((x, y)\)) provides \((\epsilon, \delta)\) membership privacy to \(z\) if the expected loss of the models not trained on \(z\) is \(\epsilon\)-close to the loss of \(\theta\) on \(z\), with probability at least \(1 - \delta\). They assume a posterior distribution of the parameters trained on a given data \(D = \{z_1, ..., z_n\}\) to be:

\[
\mathbb{P}(\theta | z_1, ..., z_n) \propto \exp \left( \sum_{i=1}^{n} \ell(\theta, z_i) \right)
\]

Consider a neighboring dataset \(D' = \{z_1, ..., z'_i, ..., z_n\}\) of \(D\), which is obtained by modifying at most one sample of \(D\) (Ding et al. 2018). Sablayrolles et al. (2019) show that, to provide membership privacy to \(z_j\), the log of the ratio of probabilities of obtaining the same \(\theta\) from \(D\) and \(D'\) should be bounded, i.e., (4) should be bounded.

\[
\log \left| \frac{\mathbb{P}(\theta | D)}{\mathbb{P}(\theta | D')} \right| = |\ell(\theta, z_j) - \ell(\theta, z'_j)|
\]

(4) implies that, if \(\theta\) was indeed trained on \(z_j\), then to provide membership privacy to \(z_j\), the loss of \(\theta\) on \(z_j\) should be same as the loss on any non-member sample \(z'_j\).

**DMP is a strong meta-regularization technique built on this intuition.** DMP aims to protect its target models against the membership inference attacks that exploit the gap between the target model’s losses on the members and non-members, by reducing the gap.

DMP achieves this via knowledge transfer and restricts the direct access of \(\theta_{up}\) to the private \(D_{tr}\), which significantly reduces the membership information leakage to \(\theta_{up}\). However, unlike existing knowledge transfer, DMP proposes an entropy-based criterion to select/generate \(X_{ref}\). Simply put, soft labels of the unprotected model \(\theta_{up}\) on \(X_{ref}\) should have low entropy and the \(X_{ref}\) should be far from decision boundaries of \(\theta_{up}\), i.e., far from \(D_{tr}\), in the input feature space. Intuitively, such samples are easy to classify and none of the members of \(D_{tr}\) significantly affects their predictions, and therefore, these predictions do not leak membership information of any particular member.

**Details of the DMP Technique** We now detail the three phases of our DMP defense depicted in Figure 1. In pre-distillation phase (step (1) in Figure 1), DMP trains \(\theta_{up}\) on the private training data, \(D_{tr}\), using standard SGD optimizer, e.g., Adam. Such unprotected \(\theta_{up}\) is highly susceptible to MIA due to large generalization error, i.e., difference between train and test accuracies (Shokri et al. 2017; Yeom et al. 2018).

Next, in distillation phase (step (2.1) in Figure 1), DMP obtains \(X_{ref}\) required to transfer the knowledge of \(\theta_{up}\) in \(\theta_p\). Note that, \(X_{ref}\) is unlabeled and cannot be used directly
Figure 1: Distillation for Membership Privacy (DMP) defense. (1) In pre-distillation phase, DMP trains an unprotected model $\theta_{up}$ on the private training data without any privacy protection. (2.1) In distillation phase, DMP uses $\theta_{up}$ to select/generate appropriate reference data $X_{ref}$ that minimizes membership privacy leakage. (2.2) Then, DMP transfers the knowledge of $\theta_{up}$ by computing predictions of $\theta_{up}$ on $X_{ref}$, denoted by $\theta_{up}^{X_{ref}}$. (3) In post-distillation phase, DMP trains the final protected model $\theta_p$ on $(X_{ref}, \theta_{up}^{X_{ref}})$. for any learning. Then, we compute soft labels of $X_{ref}$, i.e., $\theta_{up}^{X_{ref}} = \theta_{up}(X_{ref})$ (step (2.2) in Figure 1). There are two key factors of the distillation phase that allow us to tune DMP and achieve the desired privacy-utility tradeoffs. First, the lower the entropy of predictions $\theta_{up}^{X_{ref}}$, the lower the membership leakage through $X_{ref}$ and vice-versa. Second, using higher softmax temperatures to compute $\theta_{up}^{X_{ref}}$ reduces membership leakage, but may reduce accuracy of the final model, and vice-versa.

Finally, in Post-distillation phase (step (3) in Figure 1), DMP trains a protected model $\theta_p$ on $(X_{ref}, \theta_{up}^{X_{ref}})$ using Kullback-Leibler divergence loss defined in (5). In (5), $\overline{y}_i$ is the target soft label. The final $\theta_p$ is obtained by solving (6).

$$L_{KL}(\mathbf{x}, \overline{y}_i) = \sum_{i=0}^{c-1} \frac{1}{N_{ref}} \sum_{(\mathbf{x}, \overline{y}_i) \in (X_{ref}, \theta_{up}^{X_{ref}})} \log \left( \frac{\overline{y}_i}{\theta_p(\mathbf{x})} \right)$$

$$\theta_p = \arg\min_{\theta} \frac{1}{N_{ref}} \sum_{(\mathbf{x}, \overline{y}_i) \in (X_{ref}, \theta_{up}^{X_{ref}})} L_{KL}(\mathbf{x}, \overline{y}_i)$$

Due to KL-divergence loss in (6), the resulting model, $\theta_p$, perfectly learns the behavior of $\theta_{up}$ on the $X_{ref}$. Furthermore, $X_{ref}$ being a representative non-member data, i.e., test data, we expect that the test accuracies of $\theta_p$ and $\theta_{up}$ are close, and that the final DMP models will not suffer significant accuracy reductions (Ba et al. 2014,Romero et al. 2014).

**Fine-tuning the DMP Defense** As mentioned before, the appropriate choice of reference data $X_{ref}$ is important to achieve the desired privacy-utility tradeoffs in DMP. In this section, we show that $X_{ref}$ with low entropy predictions of unprotected model $\theta_{up}$ strengthens membership privacy and derive an entropy-based criterion to select/generate $X_{ref}$.

**Proposition 1.** Consider $\theta_{up}$ trained on a private $D_{tr}$. Then, the membership leakage about $D_{tr}$ through predictions $\theta_{up}(X_{ref})$ can be reduced by selecting/generating $X_{ref}$ that are far from $D_{tr}$ in input feature space with respect to some $L_p$ distance and whose predictions, $\theta_{up}(X_{ref})$, have low entropies.

**Sketch of proof of Proposition 1.** Due to space limitations, we defer the detailed proof to the full version (Shejwalkar et al. 2019) and provide its sketch here. Consider two training datasets $D_{tr}$ and $D'_{tr}$ such that $D'_{tr} \leftarrow D_{tr} - \epsilon$, and $X_{ref}$. Then, the log of the ratio of the posterior probabilities of learning the exact same parameters $\theta_p$ using DMP is given by (7). Observe that, $R$ is an extension of (4) to the setting of DMP, where $\theta_p$ is trained via the knowledge transferred using $(X_{ref}, \theta_{up}^{X_{ref}})$, instead of directly training on $D_{tr}$. Sablayrolles et al. (2019) argue to reduce this ratio to improve membership privacy. Hence, we want to obtain $X_{ref}$ which reduces $R$ when $D_{tr}$, $D'_{tr}$, and $\theta_p$ are kept constant. We note that, although similar in appearance to differential privacy, $R$ is defined only for the given private dataset, $D_{tr}$.

$$R = \left| \log \left( \Pr(\theta_p | D_{tr}, X_{ref}) / \Pr(\theta_p | D'_{tr}, X_{ref}) \right) \right|$$

Next, we modify $R$ as:

$$R = \left| -\frac{1}{N_{ref}} \sum_{x \in X_{ref}} L_{KL}((x, \theta_{up}^{x}); \theta_p) - L_{KL}((x, \theta_{up}^{x}); \theta_p) \right|$$

$$\leq \frac{1}{N_{ref}} \sum_{x \in X_{ref}} \left| L_{KL}((\theta_{up}^{x}); \theta_p) - L_{KL}((\theta_{up}^{x}); \theta_p) \right|$$

where $\theta_{up}$ and $\theta_p$ are trained on $D_{tr}$ and $D'_{tr}$, respectively. Note that, (8) holds due to the assumption in (3) and the KL-divergence loss used to train $\theta_p$ in DMP. (9) follows from (8) because $|a|+|b| \leq |a+b|$. Therefore, minimizing (9) implies minimizing (7). Thus, to improve membership privacy due to $\theta_p$, $X_{ref}$ is obtained by solving (10).

$$X_{ref} = \arg\min_{X_{ref} \in X} \frac{1}{N_{ref}} \sum_{x \in X_{ref}} \left| L_{KL}((\theta_{up}^{x}); \theta_p) - L_{KL}((\theta_{up}^{x}); \theta_p) \right|$$

The objective of (10) is minimized when $\theta_{up}^{X_{ref}} = \theta_{up}$ $\forall x \in X_{ref}$ and is very intuitive: It implies that, $z$ (i.e., $D_{tr} - D'_{tr}$) enjoys stronger membership privacy when the reference data, $X_{ref}$, are such that the distributions of outputs of $\theta_{up}$ and $\theta_{up}^X$ on $X_{ref}$ are not affected by the presence of $z$ in $D_{tr}$.

Next, we simplify (10) by replacing $L_{KL}$ with closely related cross-entropy loss $L_{CE}$. This simplification can be easily validated using $X_{ref}$ whose ground truth labels are known. Specifically, we randomly sample $D_{tr}$ and $X_{ref}$ from Purchase100 dataset, and compute $\theta_{up}$ and $\theta_p$ using DMP. Next, for some $z \in D_{tr}$, we train $\theta_{up}$ on $D'_{tr}$. Then, for each $x \in X_{ref}$, we compute $\Delta L_{CE}$ as in (10) and use the available ground truth label of $x$ to compute $\Delta L_{CE}$. Finally, we show that $\Delta L_{KL}$ and $\Delta L_{CE}$ are strongly correlated for all $z \in D_{tr}$.

Next, we use the linear approximation given by (Koh and Liang 2017) for the difference in $L_{CE}$ of a pair of models.
Empirical Verification Of Proposition 1. We randomly pick $D_{tr}$ of size 10k from Purchase100 data and train $\theta_{up}$. Then, we sort the rest of Purchase100 data based on entropy of the predictions of $\theta_{up}$ on the data. We form first $X_{ref}$ using the first 10k data with the lowest entropies, second $X_{ref}$ using the following 10k data, and so on. Finally we train multiple protected models, $\theta_{p}$’s, using each of the $X_{ref}$’s. Figure 2 (left) shows the increase in the MIA risk and Figure 2 (right) shows the increase in the classification performance of $\theta_{p}$ with the increase in average entropy of the $X_{ref}$ used. This tradeoff is because, although the higher entropy predictions contain more useful information (Nayak et al. 2019; Hinton, Vinyals, and Dean 2014) and lead to high accuracy of $\theta_{p}$, they also contain higher membership information about $D_{tr}$ and lead to higher MIA risk.

Experimental Setup

Datasets And Target Model Architectures

We use four datasets and corresponding model architectures that are consistent with the previous works (Shokri et al. (2017); Nasr et al. (2019; 2018); Salem et al. (2019)).

Purchase (Purchase 2017) is a 100 class classification task with 197,324 binary feature vectors of length 600; each dimension corresponds to a product and its value states if corresponding customer purchased the product; the corresponding label represents the shopping habit of the customer.

Texas (Texas 2017) is dataset of patient records. It is a 100 class classification task with 67,300 binary feature vectors of length 6,170 each; each dimension corresponds to symptoms and its value states if corresponding patient has the symptom or not; the label represents the treatment given to the patient. For Purchase and Texas we use fully connected (FC) networks.

CIFAR10 and CIFAR100 are popular image classification datasets, each has size 50k and $32 \times 32$ color images. We use Alexnet, DenseNet-12 (with 0.77M parameters), and DenseNet-19 (with 25.6M parameters) models for CIFAR100, and Alexnet for CIFAR10. Following previous works, we measure the test accuracy of the target models as their utility.

Sizes Of Dataset Splits. The dataset splits are given in Table 1. For Purchase and Texas tasks, we use $D_{ref}$ of size 10k and select $X_{ref}$ of size 10k from the remaining data using our entropy-based criterion. For CIFAR datasets, we use $D_{ref}$ of size 25k and due to small sizes of these datasets, use the entire remaining 25k data as $X_{ref}$. The ‘Attack training’ (described shortly) column shows the MIA adversary’s knowledge of members and non-members of $D_{tr}$. Following all the previous works, we assume that the adversary knows 50% of $D_{tr}$. Further experimental details are provided in Appendix.

| Dataset     | $|D_{tr}|$ | $|X_{ref}|$ | $|D_{p}|$ | $|D'_{tr}|$ |
|-------------|----------|-----------|---------|---------|
| Purchase (P)| 10000    | 10000     | 5000    | 5000    |
| Texas (T)   | 10000    | 10000     | 5000    | 5000    |
| CIFAR100 (C100) | 25000   | 25000     | 12500   | 8000    |
| CIFAR10 (C10)   | 25000   | 25000     | 12500   | 8000    |

Table 1: All the dataset splits are disjoint. $D$, $D'$ data are the members and non-members of $D_{tr}$ known to MIA adversary.

Membership Inference Attacks

We briefly review the four MIAAs we use for evaluations. Following previous works, we use the accuracy of MIAs on target models as a measure of their membership privacy.

Bounded loss (BL) attack (Yeom et al. 2018) decides membership using a threshold on the target model’s loss on the target sample. When 0-1 loss is used, the attack accuracy is simply the difference in training and test accuracy of target model. We denote BL attack accuracy by $A_{bl}$.

NN attack (Salem et al. 2019) uses a shadow dataset $d_s$ drawn from the same distribution as $D_{tr}$. The attacker splits $d_s$ in $d_s'$ and $d_s''$, trains a shadow model $\theta_{s}$ on $d_s'$, computes predictions of $\theta_{s}$ on $d_s'$ and $d_s''$, labels the predictions of $d_s'$ as members and that of $d_s''$ as non-members, and trains binary attack model on the predictions. We denote NN attack accuracy by $A_{nn}$. Due to their small sizes, DMP cannot be evaluated with CIFAR datasets, hence we omit NN attack evaluation for CIFAR datasets.

NSH attacks (Nasr et al. 2019) are similar to NN attacks. They concatenate various whitebox (e.g., model gradients) and/or blackbox (e.g., model loss, predictions) features of target model, while NN attack uses only the target model predictions. We denote whitebox and blackbox NSH attack accuracies by $A_{wb}$ and $A_{bb}$, respectively. For NN and NSH attacks, we use the same attack models as the original works.

Experiments

Comparison With Regularization Techniques

Regularization improves the generalization of ML models, and hence, reduce the MIA risk (Shokri et al. 2017). Hence, we compare DMP with five regularization defenses, including state-of-the-art MIA defense—adversarial regularization (Nasr et al. 2018). In all tables, $E_{gen}$ is generalization error, i.e., $(A_{train} - A_{test})$, where $A_{train}$ and $A_{test}$ are train and test accuracies of the target model, respectively. $A_{test}$ gives the % increase in $A_{test}$ due to DMP over the other regularizers. $A_{wb}$, $A_{bb}$, $A_{bi}$, $A_{nn}$ are accuracies of various attacks discussed in the previous section.
Table 2: Models trained without any defenses have high test accuracies, $A_{\text{test}}$, but their high generalization errors, $E_{\text{gen}}$ (i.e., $A_{\text{train}} - A_{\text{test}}$) facilitate strong MIAs ($\frac{\epsilon}{\Delta}$). “N/A” means the attack is not evaluated due to lack of data.

Table 2 shows accuracies of models trained without any defense; CIFAR models have lower than state-of-the-art accuracies due to smaller training datasets.

Comparison With Adversarial Regularization (AdvReg)

Table 3 compares $A_{\text{test}}$ of DMP and AdvReg models, for similar MIA accuracies (i.e., membership privacy). As expected, these models also have similar $E_{\text{gen}}$'s. However, $A_{\text{test}}$ of DMP models is significantly higher than AdvReg models; $A_{\text{test}}^+$ column shows the % increase in $A_{\text{test}}$ due to DMP over AdvReg: Accuracy improvements due to DMP over AdvReg are close to 100% for CIFAR-100, and 20% to 45% for other datasets. AdvReg uses accuracy of an MIA model to regularize and train its target models to fool the MIA model. However, AdvReg allows its target models to directly access $D_{\text{tr}}$. Hence, to effectively fool the MIA model, it puts relatively large weight on the regularization-loss term. This reduces the impact of the loss on main task and reduces the accuracy of AdvReg models. DMP uses appropriate reference data to transfer the knowledge of $D_{\text{tr}}$ to its target models without allowing them direct access. Hence, DMP significantly outperforms AdvReg in terms of privacy-utility tradeoffs.

Comparison With Other Regularizers

Next, we compare DMP with four state-of-the-art regularizers: weight decay (WD), dropout (Srivastava et al. 2014) (DR), label smoothing (Szegedy et al. 2016) (LS), and confidence penalty (Pereyra et al. 2017) (CP). Due to the poor MIA resistance of CP, we defer its results to Appendix. Table 4 shows the results, when MIA risks of regularized models is close that of DMP models (Table 3). We note that, in all the cases, $A_{\text{test}}$ of DMP are significantly higher (up to 385% increase as $A_{\text{test}}^+$ column specifies) than $A_{\text{test}}$ of other regularizers. This is because, these regularizers aim to improve the test accuracies of target models, but are not designed to reduce MIA risk. Thus, to reduce MIA risk, these regularization techniques add large, suboptimal noises during training, and hurt the utility of resulting models.

Comparison With Differentially Private Defenses

Comparison With DP-SGD

Following the methodology of Jayaraman et al. (2019), we compare DMP and DP-SGD (Abadi et al. 2016) using the empirically observed tradeoffs between membership privacy (MIA resistance) and $A_{\text{test}}$ of models. We use only CIFAR10 for these experiments, as the DP-SGD achieves prohibitively low accuracies on difficult tasks such as Texas and CIFAR100. We evaluate MIA risk using the whitebox NSH attack. Table 5 shows the results of Alexnet trained on CIFAR10 using DMP and DP-SGD with different privacy budgets $\epsilon$'s; $-ve$ $E_{\text{gen}}$ implies $A_{\text{train}}$ is lower than $A_{\text{test}}$. DP-SGD incurs significant (35%) loss in $A_{\text{test}}$ at lower $\epsilon$ (12.5) to provide strong membership privacy. At higher $\epsilon$, $A_{\text{test}}$ of DP-SGD increases, but at the cost of very high generalization error, which facilitates stronger MIAs. Note that, further increase in privacy budget, $\epsilon$, does not improve tradeoff of DP-SGD. More importantly, for low MIA risk of $\sim$ 51.3%, DMP models have 12.8% higher $A_{\text{test}}$ (i.e., 24.5% improvement) than DP-SGD models, which shows the superior tradeoffs due to DMP.

Comparison With PATE

PATE (Papernot et al. 2017), a semi-supervised learning technique, requires a compatible pair of generator and discriminator to achieve acceptable performances. Hence, we use CIFAR10 dataset and, instead of Alexnet, use the generator-discriminator pair from (Salimans et al. 2016), which has state-of-the-art performances. PATE trains a set of teachers, computes hard labels of each teacher on some $X_{\text{ref}}$, aggregates the labels for each $x \in X_{\text{ref}}$ using majority voting, adds DP noise to the aggregate, and finally trains its target model on the noisy aggregate.

We train ensembles of 5, 10, and 25 teachers using $D_{\text{tr}}$ of size 25k. We use the optimized confident-GNMax (GNMax) aggregation scheme of (Papernot et al. 2018) to label $X_{\text{ref}}$. We present a subset of results in Table 6 and defer comprehensive comparison to Appendix. At low $\epsilon$'s ($<10$), GNMax hardly produces any labels, hence, the final target model has very low $A_{\text{test}}$, but at higher $\epsilon$'s ($>1000$), PATE target model has acceptable $A_{\text{test}}$. However, PATE cannot achieve performances close to DMP, as it divides $D_{\text{tr}}$ among its teachers. Such teachers have significantly lower accuracies and their ensemble cannot achieve the accuracy close to that of the unprotected model of DMP, which is trained on the entire $D_{\text{tr}}$. Hence, the quality of knowledge transferred in DMP is always higher than that in PATE.

Discussions

Below, we provide further key insights in to DMP defense and defer their detailed discussion to Appendix.

Hyperparameter Selection In DMP

Increasing the temperature of softmax layer of the unprotected model, $\theta_{\text{up}}$, used to transfer the knowledge of $\theta_{\text{up}}$, can further reduce the membership leakage of $D_{\text{tr}}$. This is because, at higher softmax temperatures, predictions of $\theta_{\text{up}}$ have uniform distribution over all classes and contain no useful information for MIAs. Similarly, reducing the size of $X_{\text{ref}}$ reduces MIA risk due to DMP, but comes at the cost of reduction in $A_{\text{test}}$.

Privacy Risk to Reference Data ($X_{\text{ref}}$)

We evaluate the privacy risk to $X_{\text{ref}}$, as it can be of sensitive nature, e.g., in case of Texas medical records dataset. Our results in appendix show that given the final DMP model, $\theta_{\text{ref}}$, and a target sample, MIA adversary (who mounts BL, NN, or NSH attacks) cannot decide if the sample belonged to $X_{\text{ref}}$ with
Table 3: Comparing test accuracy ($A_{test}$) and generalization error ($E_{gen}$) of DMP and Adversarial Regularization, for near-equal, low MIA risks (high membership privacy). $A_{test}$ shows the % increase in $A_{test}$ due to DMP over the corresponding regularizers.

Table 4: Evaluating three state-of-the-art regularizers, with similar, low MIA risks (high membership privacy) as DMP. $A_{test}$ shows the % increase in $A_{test}$ due to DMP over the corresponding regularizers.

Table 5: DP-SGD versus DMP for CIFAR10 and Alexnet. For low MIA risk of $\sim 51.3\%$, DMP achieves $24.5\%$ higher $A_{test}$ than of DP-SGD (12.8% absolute increase in $A_{test}$).

Table 6: Comparing PATE with DMP. DMP has $E_{gen}$, $A_{test}$, and $A_{wb}$ of 1.19%, 76.79%, and 50.8%, respectively. PATE has low accuracy even at high privacy budgets, as it divides data among teachers and produces low accuracy ensembles.

Adaptive Attack On DMP In DMP, the reference data, $X_{ref}$, is selected such that the predictions of DMP’s unprotected model $\theta_{up}$ on $X_{tr}$ have low entropies. Due to memorization, predictions of $\theta_{up}$ on $D_{tr}$ also have low entropies. Hence, an adaptive adversary may exploit this peculiar $X_{ref}$ selection in DMP. Based on this intuition, we investigate the possibility of an adaptive MIA, which labels a target sample as a member if the sample is close to some $X_{ref}$ datum in feature space. However, such attack has accuracy close to random guess. This is because, we observe that the proximity of two samples in feature space has no correlation with the entropy of predictions of given $\theta_{up}$ on those samples, which is the selection criterion of DMP. We leave further investigation of adaptive attacks on DMP to future work.

Conclusions We proposed Distillation for Membership Privacy (DMP), a knowledge distillation based defense against membership inference attacks that significantly improves the membership privacy-model utility tradeoffs compared to state-of-the-art defenses. We provided a novel criterion to generate/select reference data in DMP and achieve the desired tradeoffs. Our extensive evaluation demonstrated the state-of-the-art privacy-utility tradeoffs of DMP.
Acknowledgments

This work was supported in part by NSF grant CPS-1739462.

References


Papernot, N.; Abadi, M.; Erlingsson, U.; Goodfellow, I.; and Talwar, K. 2017. Semi-supervised knowledge transfer for deep learning from private training data. International Conference on Learning and Representation.


