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 accuracies 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 ***D**istillation for **M**embership **P**rivacy* (DMP), leverages *knowledge distillation* (Hinton et al. 2014), which transfers the knowledge

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

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 ***D**istillation for **M**embership **P**rivacy* (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 trainin 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.

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 $(\mathbf{x}, y)$ in the private training data D_{tr} of a target model θ. 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 θ and a sample $(\mathbf{x}, y)$. MIAs exploit the differences in the behavior of θ 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 θ for given $(\mathbf{x}, y)$, e.g., prediction $\theta(\mathbf{x}, y)$, θ's loss on $(\mathbf{x}, y)$, and the gradients of the loss. The adversary combines these features to form $F(\mathbf{x}, y, \theta)$. The attack model h takes $F(\mathbf{x}, y, \theta)$ as its input and outputs the probability that $(\mathbf{x}, y)$ is a member of D_{tr}. Let $\Pr_{D_{\mathsf{tr}}}$ and $\Pr_{\setminus D_{\mathsf{tr}}}$ 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:

$$G^{\theta}(h) = \underset{\substack{(\mathbf{x},y)\\ \sim \Pr_{D_{\mathsf{tr}}}}}{\mathbb{E}}[\log(h(F))] + \underset{\substack{(\mathbf{x},y)\\ \sim \Pr_{\setminus D_{\mathsf{tr}}}}}{\mathbb{E}}[\log(1-h(F))] \quad (1)$$

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

$$G^{\theta}_{D^A, D'^A}(h) = \sum_{\substack{(\mathbf{x},y)\\ \in D^A}} \frac{\log(h(F))}{|D^A|} + \sum_{\substack{(\mathbf{x},y)\\ \in D'^A}} \frac{\log(1-h(F))}{|D'^A|} \quad (2)$$

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 θ_{up}. An ML model is called *protected* model, denoted by θ_{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 θ on $\mathbf{x}$, i.e., $\theta(\mathbf{x})$, by $\theta^{\mathbf{x}}$.

Main Intuition Of DMP Sablayrolles et al. (2019) show that θ trained on a sample z (short for $(\mathbf{x}, y)$) provides (ϵ, δ) membership privacy to z if the expected loss of the models not trained on z is ϵ-close to the loss of θ 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:

$$\mathbb{P}(\theta | z_1, ..., z_n) \propto \exp(\sum_{i=1}^{n} \ell(\theta, z_i)) \quad (3)$$

Consider a neighboring dataset $D' = \{z_1, .., z'_j, .., z_n\}$ 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 θ from D and D' should be bounded, i.e., (4) should be bounded.

$$\log\Big|\frac{\mathbb{P}(\theta|D)}{\mathbb{P}(\theta|D')}\Big| = |\ell(\theta, z_j) - \ell(\theta, z'_j)| \quad (4)$$

(4) implies that, if θ was indeed trained on z_j, then to provide membership privacy to z_j, the loss of θ 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 θ_{p} to the private D_{tr}, which significantly reduces the membership information leakage to θ_{p}. However, unlike existing knowledge transfer, DMP proposes an entropy-based criterion to select/generate X_{ref}. Simply put, soft labels of the unprotected model θ_{up} on X_{ref} should have low entropy and the X_{ref} should be far from decision boundaries of θ_{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 θ_{up} on the private training data, D_{tr}, using standard SGD optimizer, e.g., Adam. Such unprotected θ_{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 θ_{up} in θ_{p}. Note that, X_{ref} is *unlabeled* and cannot be used directly

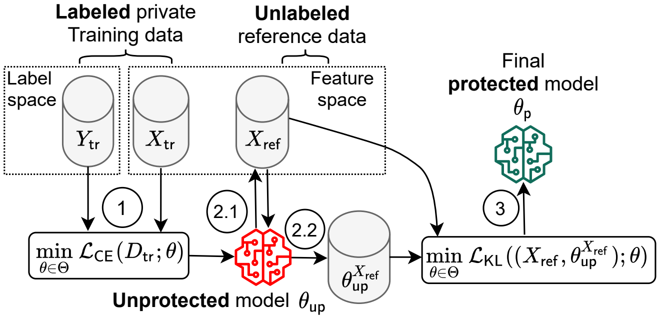
Labeled private
Training data
Unlabeled
reference data
Label
space
Feature
space
Y_{tr}
X_{tr}
X_{ref}
Final
protected model
θ_{p}
1
2.1
2.2
3
$\min_{\theta\in\Theta} \mathcal{L}_{\text{CE}}(D_{\text{tr}};\theta)$
$\theta_{\text{up}}^{X_{\text{ref}}}$
$\min_{\theta\in\Theta} \mathcal{L}_{\text{KL}}((X_{\text{ref}},\theta_{\text{up}}^{X_{\text{ref}}});\theta)$
Unprotected model θ_{up}

Figure 1: Distillation for Membership Privacy (DMP) defense. (1) In pre-distillation phase, DMP trains an unprotected model θ_{up} on the private training data without any privacy protection. (2.1) In distillation phase, DMP uses θ_{up} to select/generate appropriate reference data X_{ref} that minimizes membership privacy leakage. (2.2) Then, DMP transfers the knowledge of θ_{up} by computing predictions of θ_{up} on X_{ref}, denoted by $\theta_{\text{up}}^{X_{\text{ref}}}$. (3) In post-distillation phase, DMP trains the final protected model θ_{p} on $(X_{\text{ref}}, \theta_{\text{up}}^{X_{\text{ref}}})$.

for any learning. Then, we compute soft labels of X_{ref}, i.e., $\theta_{\text{up}}^{X_{\text{ref}}} = \theta_{\text{up}}(X_{\text{ref}})$ (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 $\theta_{\text{up}}^{X_{\text{ref}}}$, the lower the membership leakage through X_{ref} and vice-versa. Such low entropy predictions are characteristics of the members of D_{tr}, however, non-members with low entropy can be obtained (or generated using GANs (Micaelli and Storkey 2019)) due to large input feature space. Second, using higher softmax temperatures to compute $\theta_{\text{up}}^{X_{\text{ref}}}$ 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 θ_{p} on $(X_{\text{ref}}, \theta_{\text{up}}^{X_{\text{ref}}})$ using Kullback-Leibler divergence loss defined in (5). In (5), $\bar{\mathbf{y}}$ is the target soft label. The final θ_{p} is obtained by solving (6).

$$\mathcal{L}_{\text{KL}}(\mathbf{x},\bar{\mathbf{y}}) = \sum_{i=0}^{\mathbf{c}-1} \bar{\mathbf{y}}_i \log\left(\frac{\bar{\mathbf{y}}_i}{\theta_{\text{p}}(\mathbf{x})_i}\right) \tag{5}$$

$$\theta_{\text{p}} = \underset{\theta}{\operatorname{argmin}} \frac{1}{|X_{\text{ref}}|} \sum_{(\mathbf{x},\bar{\mathbf{y}})\in(X_{\text{ref}},\theta_{\text{up}}^{X_{\text{ref}}})} \mathcal{L}_{\text{KL}}(\mathbf{x},\bar{\mathbf{y}}) \tag{6}$$

Due to KL-divergence loss in (6), the resulting model, θ_{p}, perfectly learns the behavior of θ_{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 θ_{p} and θ_{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 θ_{up} strengthens membership privacy and derive an entropy-based criterion to select/generate X_{ref}.

Proposition 1. *Consider θ_{up} trained on a private D_{tr}. Then, the membership leakage about D_{tr} through predictions $\theta_{\text{up}}(X_{\text{ref}})$ 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, $\theta_{\text{up}}(X_{\text{ref}})$, 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 $D'_{\text{tr}} \leftarrow D_{\text{tr}} - z$, and X_{ref}. Then, the log of the ratio of the posterior probabilities of learning the exact same parameters θ_{p} using DMP is given by (7). Observe that, $\mathcal{R}$ is an extension of (4) to the setting of DMP, where θ_{p} is trained via the knowledge transferred using $(X_{\text{ref}}, \theta_{\text{up}}^{X_{\text{ref}}})$, 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 $\mathcal{R}$ when D_{tr}, D'_{tr}, and θ_{p} are kept constant. We note that, although similar in appearance to differential privacy, $\mathcal{R}$ is defined only for the given private dataset, D_{tr}.

$$\mathcal{R} = \left|\log\left(\Pr(\theta_{\text{p}}|D_{\text{tr}}, X_{\text{ref}})/\Pr(\theta_{\text{p}}|D'_{\text{tr}}, X_{\text{ref}})\right)\right| \tag{7}$$

Next, we modify $\mathcal{R}$ as:

$$\mathcal{R} = \Big| -\frac{1}{T}\sum_{\mathbf{x}\in X_{\text{ref}}} \mathcal{L}_{\text{KL}}((\mathbf{x},\theta_{\text{up}}^{\mathbf{x}});\theta_{\text{p}}) - \mathcal{L}_{\text{KL}}((\mathbf{x},\theta'^{\mathbf{x}}_{\text{up}});\theta_{\text{p}})\Big| \tag{8}$$

$$\leq \frac{1}{T}\sum_{\mathbf{x}\in X_{\text{ref}}} \Big|\mathcal{L}_{\text{KL}}(\theta_{\text{up}}^{\mathbf{x}}\|\theta_{\text{p}}^{\mathbf{x}}) - \mathcal{L}_{\text{KL}}(\theta'^{\mathbf{x}}_{\text{up}}\|\theta_{\text{p}}^{\mathbf{x}})\Big| \tag{9}$$

where θ_{up} and θ'_{up} 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 θ_{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 θ_{p}, X_{ref} is obtained by solving (10).

$$X^*_{\text{ref}} = \underset{X_{\text{ref}}\in X}{\operatorname{argmin}}\Big(\frac{1}{T}\sum_{\mathbf{x}\in X_{\text{ref}}} \big|\mathcal{L}_{\text{KL}}(\theta_{\text{up}}^{\mathbf{x}}\|\theta_{\text{p}}^{\mathbf{x}}) - \mathcal{L}_{\text{KL}}(\theta'^{\mathbf{x}}_{\text{up}}\|\theta_{\text{p}}^{\mathbf{x}})\big|\Big) \tag{10}$$

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

Next, we simplify (10) by replacing $\mathcal{L}_{\text{KL}}$ with closely related cross-entropy loss $\mathcal{L}_{\text{CE}}$. 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 θ_{up} and θ_{p} using DMP. Next, for some $z \in D_{\text{tr}}$, we train θ'_{up} on D'_{tr}. Then, for each $\mathbf{x} \in X_{\text{ref}}$, we compute $\Delta\mathcal{L}_{\text{KL}}$ as in (10) and use the available ground truth label of $\mathbf{x}$ to compute $\Delta\mathcal{L}_{\text{CE}}$. Finally, we show that $\Delta\mathcal{L}_{\text{KL}}$ and $\Delta\mathcal{L}_{\text{CE}}$ are strongly correlated for all $z \in D_{\text{tr}}$.

Next, we use the linear approximation given by (Koh and Liang 2017) for the difference in $\mathcal{L}_{\text{CE}}$ of a pair of models

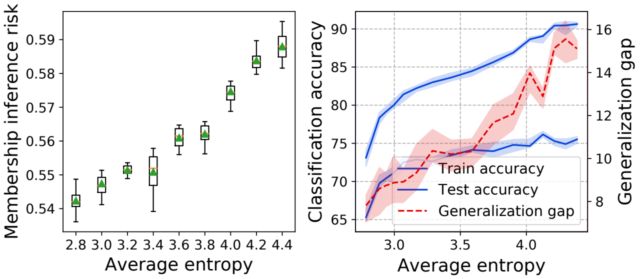

Figure 2: The lower the entropy of predictions of unprotected model on X_{ref}, the higher the membership privacy.

trained with and without a sample to simplify (10). Then the result of Proposition 1 follows after a few simple mathematical manipulations.

Empirical Verification Of Proposition 1. We randomly pick D_{tr} of size 10k from Purhcase100 data and train θ_{up}. Then, we sort the rest of Purhcase100 data based on entropy of the predictions of θ_{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, θ_{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 θ_{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 θ_{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 × 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	DMP training		Attack training	
	$\lvert D_{\text{tr}}\rvert$	$\lvert X_{\text{ref}}\rvert$	$\lvert D\rvert$	$\lvert D'\rvert$
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 MIAs 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* θ_s on d'_s, computes predictions of θ_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 defeses, including state-of-the-art MIA defense—adeversarial regularization (Nasr et al. 2018). In all tables, E_{gen} is generalization error, i.e., $(A_{\text{train}} - A_{\text{test}})$, 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_{bl}, A_{nn} are accuracies of various attacks discussed in the previous section.

Dataset and model	No defense					
	E_{gen}	A_{test}	A_{wb}	A_{bb}	A_{bl}	A_{nn}
P-FC	24.0	76.0	77.1	76.8	63.1	60.5
T-FC	51.3	48.7	84.0	82.2	76.1	71.9
C100-A	63.2	36.8	90.3	91.3	81.8	N/A
C100-D12	33.8	65.2	72.2	71.8	67.5	N/A
C100-D19	34.4	65.5	82.3	81.6	68.1	N/A
C10-A	32.5	67.5	77.9	77.5	66.4	N/A

Table 2: Models trained without any defenses have high test accuracies, A_{test}, but their high generalization errors, E_{gen} (i.e., $A_{\mathsf{train}} - A_{\mathsf{test}}$) facilitate strong MIAs (§). "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_{test} of DMP and AdvReg models, for similar MIA accuracies (i.e., membership privacy). As expected, these models also have similar E_{gen}'s. However, A_{test} of DMP models is significantly higher than AdvReg models; A^{+}_{test} column shows the % increase in A_{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_{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_{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_{test} of DMP are significantly higher (up to 385% increase as A^{+}_{test} column specifies) than A_{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_{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 ϵ's; -ve E_{gen} implies A_{train} is lower than A_{test}. DP-SGD incurs significant (35%) loss in A_{test} at lower ϵ (12.5) to provide strong membership privacy. At higher ϵ, A_{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, ϵ, does not improve tradeoff of DP-SGD. More importantly, for low MIA risk of $\sim$ 51.3%, DMP models have 12.8% higher A_{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-supervized learning technique, requires a compatible pair of generator and disciminator 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_{ref}, aggregates the labels for each $\mathbf{x} \in X_{\mathsf{ref}}$ 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_{tr} of sise 25k. We use the optimized confident-GNMax (GNMax) aggregation scheme of (Papernot et al. 2018) to label X_{ref} We present a subset of results in Table 6 and defer comprehensive comparison to Appendix. At low ϵ's (<10), GNMax hardly produces any labels, hence, the final target model has very low A_{test}, but at higher ϵ's (>1000), PATE target model has acceptable A_{test}. However, PATE cannot achieve performances close to DMP, as it divides D_{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_{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, θ_{up}, used to transfer the knowledge of θ_{up}, can further reduce the membership leakage of D_{tr}. This is because, at higher softmax temperatures, predictions of θ_{up} have uniform distribution over all classes and contain no useful information for MIAs. Similarly, reducing the size of X_{ref} reduces MIA risk due to DMP, but comes at the cost of reduction in A_{test}.

Privacy Risk to Reference Data (X_{ref}) We evaluate the privacy risk to X_{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, θ_{p}, and a target sample, MIA adversary (who mounts BL, NN, or NSH attacks) cannot decide if the sample belonged to X_{ref} with

Dataset and model	Adversarial regularization (AdvReg)						DMP						
	E_{gen}	A_{test}	Attack accuracy				E_{gen}	A_{test}	A^+_{test}	Attack accuracy			
			A_{wb}	A_{bb}	A_{bl}	A_{nn}				A_{wb}	A_{bb}	A_{bl}	A_{nn}
Purchase + FC	9.7	56.5	55.8	55.4	54.9	50.1	10.1	74.1	**+31.2**%	55.3	55.1	55.2	50.2
Texas + FC	6.1	33.5	58.2	57.9	54.1	50.8	7.1	48.6	**+45.1**%	55.3	55.4	53.6	50.0
CIFAR100 + Alexnet	6.9	19.7	54.3	54.0	53.5	N/A	6.5	35.7	**+81.2**%	55.7	55.6	53.3	N/A
CIFAR100 + DenseNet-12	5.5	26.5	51.4	51.3	52.8	N/A	3.6	63.1	**+138.1**%	53.7	53.0	51.8	N/A
CIFAR100 + DenseNet-19	7.2	33.9	54.2	53.4	53.6	N/A	7.3	65.3	**+92.6**%	54.7	54.4	53.7	N/A
CIFAR10 + Alexnet	4.2	53.4	51.9	51.2	52.1	N/A	3.1	65.0	**+21.7**%	51.3	50.6	51.6	N/A

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} of DMP over Adversarial Regularization.

Purchase + FC (DMP's A_{test} = 74.1)						
Regularizer	E_{gen}	A_{test}	A^+_{test}	A_{wb}	A_{bb}	A_{bl}
WD	10.3	42.5	**+74.4**%	54.9	55.4	55.2
WD + DR	9.1	42.1	**+76.0**%	56.4	56.8	54.6
WD + LS	12.3	42.0	**+76.4**%	57.2	57.0	56.2

Texas + FC (DMP's A_{test} = 48.6)						
Regularizer	E_{gen}	A_{test}	A^+_{test}	A_{wb}	A_{bb}	A_{bl}
WD	5.0	22.5	**+116**%	58.3	57.7	52.5
WD + DR	6.1	14.2	**+242**%	63.1	62.6	53.1
WD + LS	8.3	37.3	**+30**%	61.7	61.0	54.2

CIFAR100 + DenseNet-12 (DMP's A_{test} = 63.1)						
Regularizer	E_{gen}	A_{test}	A^+_{test}	A_{wb}	A_{bb}	A_{bl}
WD	4.0	26.3	**+140**%	49.9	49.7	52.0
WD + DR	3.7	32.3	**+95.4**%	51.2	51.0	51.9
WD + LS	2.7	13.0	**+385**%	51.0	51.4	51.4

CIFAR10 + Alexnet (DMP's A_{test} = 65.0)						
Regularizer	E_{gen}	A_{test}	A^+_{test}	A_{wb}	A_{bb}	A_{bl}
WD	4.1	45.9	**+41.6**%	52.4	52.5	52.1
WD + DR	3.2	44.7	**+45.4**%	51.9	51.7	51.6
WD + LS	4.8	53.2	**+22.2**%	53.8	53.0	52.4

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.

sufficient confidence. This is expected, because DMP trains its θ_{p} on noisy, soft-labels of X_{ref}, which do not contain any sensitive information about X_{ref} and its ground-truth labels, which is necessary for MIAs to succeed (Yeom et al. 2018). We provide detailed results in Appendix.

DMP With Synthetic Reference Data (X_{ref}) Following previous works (Papernot et al. 2018,2017), including the state-of-the-art MIA defense AdvReg (Nasr et al. 2018), we assume availability of X_{ref}. However, in privacy sensitive domains such as patient medical records, X_{ref} may not be available. Hence, we show that the assumption can be relaxed by using X_{ref} synthesized from private D_{tr} to train DMP models. For CIFAR10, we use DC-GAN to generate synthetic X_{ref} of sizes 12.5k, 25k, and 37.5k from D_{tr} of size 25k. We then train three DMP models and evaluate their MIA risk using whitebox NSH attack. We note that for 12.5k, 25k, and 37.5k synthetic X_{ref} samples, (E_{gen}, A_{test}, A_{wb}) of DMP are (2.1, 53.0, 50.3), (3.5, 56.8, 51.3), and (5.0, 57.5, 52.1), respectively. Note that, *DMP outperforms existing defenses even with synthetic* X_{ref} (Tables 3, 4).

Defense	Privacy budget (ϵ)	E_{gen}	A_{test}	A_{wb}
No defense	–	32.5	67.5	77.9
DMP	–	3.10	65.0	51.3
DP-SGD	198.5	3.60	52.2	51.7
	50.2	1.30	36.9	50.2
	12.5	0.30	31.7	50.0
	6.8	-1.60	29.4	49.9

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}).

# of Teachers	Queries answered	Privacy budget (ϵ)	Target model		A_{wb}
			E_{gen}	A_{test}	
5	49	195.9	31.4	33.9	49.1
	1163	11684	65.4	68.1	49.0
10	23	42.9	39.1	38.3	50.1
	1527	6535	63.9	65.2	49.8
25	108	183.5	53.8	55.7	49.0
	4933	1794.1	57.8	60.3	48.6

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 θ_{up} on X_{ref} have low entropies. Due to memorization, predictions of θ_{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 θ_{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

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM.

Ba, J.; and Caruana, R. 2014. Do deep nets really need to be deep? In *Advances in neural information processing systems*, 2654–2662.

Buciluǎ, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 535–541.

Carlini, N.; Liu, C.; Kos, J.; Erlingsson, U.; and Song, D. 2018. The secret sharer: Measuring unintended neural network memorization and extracting secrets. *arXiv preprint arXiv:1802.08232* .

Chaudhuri, K.; Monteleoni, C.; and Sarwate, A. D. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research* 12(Mar): 1069–1109.

Crowley, E. J.; Gray, G.; and Storkey, A. J. 2018. Moonshine: Distilling with cheap convolutions. In *Advances in Neural Information Processing Systems*, 2888–2898.

Ding, Z.; Wang, Y.; Wang, G.; Zhang, D.; and Kifer, D. 2018. Detecting violations of differential privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 475–489. ACM.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM.

Hinton, G.; Vinyals, O.; and Dean, J. 2014. Distilling the Knowledge in a Neural Network. *NIPS 2014 Deep Learning Workshop* .

Hitaj, B.; Ateniese, G.; and Pérez-Cruz, F. 2017. Deep models under the GAN: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM.

Jayaraman, B.; and Evans, D. 2019. Evaluating Differentially Private Machine Learning in Practice. In *USENIX Security Symposium*.

Jia, J.; Salem, A.; Backes, M.; Zhang, Y.; and Gong, N. Z. 2019. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 259–274.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1885–1894. JMLR. org.

Leino, K.; and Fredrikson, M. 2019. Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference. *arXiv preprint arXiv:1906.11798* .

Micaelli, P.; and Storkey, A. J. 2019. Zero-shot knowledge transfer via adversarial belief matching. In *Advances in Neural Information Processing Systems*, 9551–9561.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 634–646. ACM.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2019. Comprehensive Privacy Analysis of Deep Learning: Stand-alone and Federated Learning under Passive and Active White-box Inference Attacks. *Security and Privacy (SP), 2019 IEEE Symposium on* .

Nayak, G. K.; Mopuri, K. R.; Shaj, V.; Radhakrishnan, V. B.; and Chakraborty, A. 2019. Zero-Shot Knowledge Distillation in Deep Networks. In *International Conference on Machine Learning*, 4743–4751.

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* .

Papernot, N.; Song, S.; Mironov, I.; Raghunathan, A.; Talwar, K.; and Erlingsson, Ú. 2018. Scalable Private Learning with PATE. *arXiv preprint arXiv:1802.08908* .

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548* .

Purchase. 2017. Acquire Valued Shoppers Challenge. https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data. [Online; accessed 11-September-2019].

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550* .

Sablayrolles, A.; Douze, M.; Schmid, C.; Ollivier, Y.; and Jegou, H. 2019. White-box vs Black-box: Bayes Optimal Strategies for Membership Inference. In *International Conference on Machine Learning*, 5558–5567.

Salem, A.; Zhang, Y.; Humbert, M.; Fritz, M.; and Backes, M. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. *In NDSS* .

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*, 2234–2242.

Shejwalkar, V.; and Houmansadr, A. 2019. Reconciling Utility and Membership Privacy via Knowledge Distillation. *arXiv preprint arXiv:1906.06589* .

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *Security and Privacy (SP), 2017 IEEE Symposium on*.

Song, L.; and Mittal, P. 2020. Systematic Evaluation of Privacy Risks of Machine Learning Models. *arXiv preprint arXiv:2003.10595* .

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1): 1929–1958.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Texas. 2017. Texas hospital stays dataset. https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm. [Online; accessed 10-February-2020].

Yeom, S.; Giacomelli, I.; Fredrikson, M.; and Jha, S. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, 268–282. IEEE.

Zagoruyko, S.; and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928* .