24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Private Knowledge Transfer via Model Distillation with Generative Adversarial Networks

Di Gao and **Cheng Zhuo**[1]

Abstract. The deployment of deep learning applications has to address the growing privacy concerns when using private and sensitive data for training. A conventional deep learning model is prone to privacy attacks that can recover the sensitive information of individuals from either model parameters or accesses to the target model. Recently, differential privacy that offers provable privacy guarantees has been proposed to train neural networks in a privacy-preserving manner to protect training data. However, many approaches tend to provide the worst case privacy guarantees for model publishing, inevitably impairing the accuracy of the trained models. In this paper, we present a novel private knowledge transfer strategy, where the private teacher trained on sensitive data is not publicly accessible but teaches a student to be publicly released. In particular, a three-player (teacher-student-discriminator) learning framework is proposed to achieve trade-off between utility and privacy, where the student acquires the distilled knowledge from the teacher and is trained with the discriminator to generate similar outputs as the teacher. We then integrate a differential privacy protection mechanism into the learning procedure, which enables a rigorous privacy budget for the training. The framework eventually allows student to be trained with only unlabelled public data and very few epochs, and hence prevents the exposure of sensitive training data, while ensuring model utility with a modest privacy budget. The experiments on MNIST, SVHN and CIFAR-10 datasets show that our students obtain the accuracy losses $w.r.t$ teachers of 0.89%, 2.29%, 5.16%, respectively with the privacy bounds of (1.93, 10^{-5}), (5.02, 10^{-6}), (8.81, 10^{-6}). When compared with the existing works [15, 20], the proposed work can achieve 5-82% accuracy loss improvement.

1 INTRODUCTION

At the era of big data, the recent breakthroughs of computing infrastructures and neural network algorithms have facilitated the adoption of deep learning in various domains from facial recognition to health care management. The successful deep learning applications in real-world services depend on not only the high-performance inference models but also the quantity and the quality of training data.

Such training data often contains private and sensitive information, $e.g.$, facial features, financial records, health history, $etc.$, inevitably causing the risk of privacy leakage for the data owners. Thus, there have been increasing privacy concerns with the growing deployment of deep learning applications. Since deep neural network itself can function as an encoder by translating the individual data into model parameters, many prior works [4, 18] have demonstrated the 'hacking' of sensitive information from the neural network, the performance of which can be notably improved if the attacker can repeatedly query outputs of the model even in a 'black-box' manner [19]. Thus, it is highly desired to have privacy-preserving techniques for deep learning applications that can ensure model utility while protect sensitive information.

Recent researches [1, 17, 24] have investigated privacy protection from various aspects. The concept of privacy-preserving deep learning was first proposed in [17], in which multiple private participants jointly trained a model by updating the sanitized model parameters, while the training data were kept at local. A more general approach was then proposed by [1] that applied differential privacy (DP) mechanisms to perturb the gradients of each iteration and employed a privacy accountant to track the privacy loss during training. Thus, by limiting the impact of one single data on model parameters, the privacy can hardly be reverse-engineered.

In practice, the aforementioned protections are prone to the worst case privacy guarantees assuming attackers have access to the internal model parameters, $i.e.$, very large noise can be injected at the cost of model utility degradation [1]. Thus, the applicability of such mechanisms are questionable. A more promising alternative is to only expose a learned model obtained through transfer learning instead of the private model directly learned on the sensitive data. The training procedure with access to the private model is prudently privacy-bounded. For example, recent works [15, 16] proposed a two-player (teacher-student) framework to train a student model on the differentially private aggregated outputs of an ensemble of teachers. Such privacy protection is achieved through the noisy voting by teachers but requires many teachers to compensate for the noise added to each query to ensure model utility. Although such a private transfer learning mitigates the exposure of private models associated with sensitive data, it still remains *a challenging task to balance between model utility and privacy when the training data is limited*. This is not a trivial problem and needs to resolve the following issues:

- **Model performance bottleneck**. Differential privacy enforces a certain level of privacy budget given the composition theorem, inevitably restricting the number of training epochs that makes it hard to further improve model performance.
- **Availability of training resources**. For privacy purpose, it is desired to have only public data fed to all the networks. However, it is difficult for the student ($e.g.$, local reservoir of a hospital) to collect sufficient quality data for training ($e.g.$, lack of labels, limited quantity, $etc.$).
- **Overhead of noisy queries**. The standard approach to DP is to inject noise to the output while the noise scale is proportional to the sensitivity of the query function. A higher sensitivity inevitably results in excessive noise, completely masking the knowledge to be transferred. For example, PATE proposes to reduce the sensitivity

[1] Zhejiang University, China, email: czhuo@zju.edu.cn

of voting results by ensembling 250 teacher models [15]. Such an approach brings significant overhead to computational resources and hence is not scalable to large-scale tasks.

In this paper, we propose a novel three-player (teacher-student-discriminator) framework that combines the state-of-art knowledge transfer techniques with the advanced privacy-preserving mechanisms. In the framework, the target model is treated as the $student$ in the conventional *teacher-student* learning paradigm and the generator in generative adversarial networks (GAN). The $student$ is not only trained by the $teacher$ through Knowledge Distillation (KD), but also adversarially trained with the $discriminator$ through GAN to generate similar outputs as the teacher. When with limited quality training data, which is often the case in practice, we enforce the discriminator to train the student to mimic the 'true' learned distribution of the teacher. The teacher in the framework is pre-trained on sensitive data and frozen during the training. Thus, unlike the two-way transfer learning as in [21], the knowledge transfer in our framework is unidirectional, $i.e.$ only the teacher distills the knowledge to the student without privacy breach. Then, to enforce privacy guarantee during training, a privacy protection mechanism is introduced that provides insights into the vulnerabilities of the framework and applies DP to protect the private teacher. Finally, the proposed learning strategy achieves the joint utility and privacy optimization that transfers private knowledge from a protected teacher to a public student.

The proposed learning strategy for private knowledge transfer exhibits three important advantages:

- The student network that learns the distilled knowledge with the discriminator is better optimized than the conventional teacher-student paradigm;
- Faster convergence can be achieved even with limited training epochs, while the performance is not overly bounded by the number of training instances;
- With the well-designed privacy-preserving mechanism, the framework is able to achieve excellent model utility and rigorous privacy guarantee.

The contributions of this work can be briefly summarized as below:

- We propose a three-player (teacher-discriminator-student) framework to transfer private knowledge to the student. With the discriminator, the student can accurately and efficiently mimic the teacher even with limited quality training data, which enables excellent model utility and a strong privacy guarantee.
- We integrate a differential privacy mechanism into the learning procedure that allows student to be trained with a rigorous privacy budget, in which the privacy accountant provides a theoretical basis for trade-off between utility and privacy.
- We evaluate the proposed framework on MNIST, SVHN and CIFAR-10 datasets. It is found that the proposed framework offers a good utility and privacy trade-off even with very few training epochs and unlabelled training instances. In addition, Our students achieve accuracies of 98.52%, 93.12% and 84.79% with DP bounds of (1.93, 10^{-5}) for MNIST, (5.02, 10^{-6}) for SVHN and (8.81, 10^{-6}) for CIFAR-10, respectively. Compared with the existing works, our framework consistently ensures a lower student accuracy loss ($w.r.t.$ the teacher). In particular, the students have only 0.89%, 2.29% and 5.16% accuracy losses for the three datasets, while the reported student accuracy losses from other works are 1.21%, 10.88%, and 5.40%, respectively [15, 20].

2 BACKGROUND AND RELATED WORK

Since the emergence of AlexNet, it is found that deeper neural networks require more training data to ensure convergence and robustness. In the healthcare, financial, or privacy-related domains, such training data is typically collected in a centralized manner and inevitably undergoes privacy breach risk if the trained models are exposed to the public. Thus, it is crucial to protect the privacy of training data as well as its use in deep neural networks. Recently various attacks further aggravate such privacy concerns in training and releasing deep learning models, as attackers can excessively analyze the model responses to recover the sensitive information. As a consequence, privacy protection in deep learning has been a concerning research area.

Differential privacy (DP) is a widely-adopted approach to privacy protection in deep learning models. The theoretical study [3] provides provable privacy guarantees for differential privacy that is achieved through adding noise to mask the output differences for the two different inputs. The very first proposal of deep learning with DP was presented in [1], in which the gradients in stochastic gradient descent (SGD) algorithm were perturbed and the privacy budgets were accordingly tracked using the *moment accountant*. Its successful development has promoted several following studies [11, 24] on differentially private deep learning.

Instead of directly applying DP to model training, references [15, 16] implemented a private transfer learning framework (PATE) that transferred the knowledge of an ensemble of teacher models to a student model. Intuitively, its privacy is guaranteed by training teachers on disjoint datasets and aggregating the outputs with noise. However, PATE requires a large number of teacher models to compensate for the noise injected to the individual query responses to ensure a desired trade-off between utility and privacy. In addition, the efficiency and effectiveness of PATE heavily relies on the correctness of voting results of the sample query as well as the appropriate noise added. This is actually hard in practice to select a voting label that is helpful to improve the training while does not reveal privacy when the student only has access to limited public (even unlabelled) data. Another representative work is [20], which introduced a private model compression framework with the conventional transfer learning and supervised learning techniques. However, the efficiency and applicability of the work [20] may be significantly impacted when the student only can use limited quality data.

Privacy accountant is an indispensable part to DP, which can track the accumulated privacy loss and enforce the applicable privacy policy [1]. It has been noted that the privacy loss radically comes from the number of queries responded by the private teacher, $i.e.$, the number of iterations during student training. A formal definition of differential privacy [2] is given as below :

Definition 1. *A randomized mechanism $\mathcal{M}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ε, δ)-differential privacy if for any two adjacent inputs D, $D' \in \mathcal{D}$ and for any subset of outputs $\mathcal{S} \subseteq \mathcal{R}$ it holds that:*

$$Pr[\mathcal{M}(D) \in \mathcal{S}] \leq e^{\varepsilon} \cdot Pr[\mathcal{M}(D' \in \mathcal{S}] + \delta. \quad (1)$$

In Eq. (1), the parameter ϵ is an upper bound for privacy loss, and parameter δ is a failure probability for this privacy guarantee. A smaller privacy bound ϵ enforces a stronger privacy guarantee but inevitably incurs more significant accuracy loss. Based on the composition theorem of DP, [1] proposed the concept of *moment accountant* to quantitatively capture the privacy budget. The extended work in [10, 13] further proposed the application of Rényi differential privacy

(RDP) accountant by analyzing Rényi divergence to enable tighter bounds of privacy loss. In this paper, we employ *RDP accountant* to measure the privacy budget of the proposed training procedure, and balance the trade-off between utility and privacy.

It is noted that direct applications of DP to the exposed models may incur too much utility loss. Other than transfer learning via an ensemble of teachers, **Knowledge distillation** (KD) is considered as an effective alternative to transfer knowledge from a private teacher to a student network to protect the privacy of training data. In the teacher-student paradigm, the teacher teaches the student through a softened distribution of its output, $a.k.a.$, dark knowledge [22]. It is easier to mimic the teacher than directly learning the target function, as the output distribution from the teacher embodies additional information beyond the ground-truth distribution. A recent study in [8] provided theoretical insights that an intelligent teacher can transfer helpful privileged information only available at the teacher's training stage. Thus, the performance of the published student model is actually associated with the distilled knowledge from the teacher network trained on private data. Such findings motivate us to investigate how to design knowledge transfer to balance between utility and privacy.

Generative adversarial network (GAN) is another alternative to enable the model to learn the true data distribution [5, 23, 25]. GAN may employ a generator to synthesize student features and a discriminator to distinguish student outputs from teacher outputs. Recently GAN has been successfully applied to generate discrete data ($e.g.$, sequence, text) [23, 25]. This is consistent with our purpose of generating/learning a discrete classification distribution for the student.

Thus, in this paper, we explore the idea of integrating KD and GAN, where *a discriminator trains the student to learn the distribution over the pseudo labels created by the teacher while the teacher distills dark knowledge to the student.* An earlier work of [21] also proposed to combine KD with GAN to help speed up the training procedure, where both the student and the teacher are trained against the discriminator to learn the true distribution and then distill knowledge to each other. Such two-way transfer learning inevitably undergoes the risk of privacy breach and hence cannot be simply applied to the private knowledge transfer.

3 OVERVIEW

This paper proposes a private knowledge transfer strategy, where the private teacher trained on sensitive data is not publicly accessible but can be used to teach a student model to be released. The target student is not only taught by the distilled knowledge from the teacher, but also trained against a discriminator to mimic the behavior of the teacher. There are two critical questions to be answered to implement the aforementioned strategy: (1) How to combine KD with GAN to train the target student model? (2) How the teacher response is privacy-guaranteed during the training procedure?

An overview of the proposed framework to implement such a strategy is demonstrated in Figure 1. The framework involves three players (teacher-student-discriminator) and two domains (non-public and public). There are two *knowledge transfer paths* for the target student. KD acts as a unidirectional path from teacher to student, and GAN is a two-way path between discriminator and student:

- **Distillation learning by KD** (detailed in Sec. 4.1): For a given input instance, we adopt the Kullback Leibler (KL) divergence as the distillation loss to measure the distance between two categorical distributions of teacher and student, $i.e.$, class probabilities from teacher and student. The effectiveness of distillation stems from the additional supervision and regularization of higher entropy soft targets.
- **Adversarial learning by GAN** (detailed in Sec. 4.2): Student (generator) and discriminator play a min-max game between one another. The discriminator tries to distinguish the student's output from the teacher's, while the student tries to generate similar output as the teacher that makes the discriminator can no longer differentiate. Both student and discriminator are adversarially trained epoch by epoch until the equilibrium.

Due to the lack of high quality training data for the student, it is typically challenging for the student to use the cross-entropy error as the objective function, which makes the training without an effective supervision. However, the combination of knowledge distillation and adversarial learning results in more effective optimization that resolves the issue. As discussed in [21], a weighted sum of the distillation loss (from the teacher) and the adversarial loss (from the discriminator) may reduce the gradient variance, thereby accelerating the student training convergence with fewer training epochs. Motivated by that, we propose a joint optimization of distillation and adversarial losses (as detailed in Sec. 4.3) to cover the *first question*, which can enforce the student to accurately mimic the teacher and ensure GAN to quickly reach the equilibrium.

To protect the privacy of sensitive data, as shown in Figure 1, we keep the pre-trained teacher model (on the sensitive data) inaccessible to the public (or attackers), $i.e.$, either internal model parameters or outputs of the model are not available to the public. The only accessible part in the framework is the student model, which can take nonsensitive public data. The connection between non-public and public domains are the two knowledge transfer paths without any direct access to private data. Since the student itself cannot access the private data, the remaining privacy protection problem for the framework is how to limit privacy loss for querying the private teacher during the student training.

Intuitively, the excessive memorization of the private teacher may expose the private and sensitive data under attacks. Thus, we propose to integrate a differential privacy (DP) mechanism into the training procedure for privacy guarantee while ensuring model utility (as detailed in Sec. 4.1), which answers the *second question* for the strategy implementation. The mechanism injects the Gaussian noise to each query to the teacher, and then track the bound of privacy loss based on the Composition Theorem of DP [2]. In other words, we sanitize the teacher's distillation loss through clipping the batch loss with the global norm bound and perturbing it with appropriate Gaussian noise. Since the discriminator in the framework is not accessible to attackers after the training, the discriminator itself and its adversarial training procedure does not incur additional privacy violations. Finally, the mechanism employs *RDP accountant* to keep track of privacy loss during training, enabling a theoretical basis for trade-off between utility and privacy.

4 Method

In this section, we formulate the proposed learning strategy with a cohort of the three networks (as plotted in Figure 2). Given a private dataset, we can always pre-train the teacher network $\mathcal{T}$ using the cross-entropy error as the objective function. On the other hand, with a nonsensitive public dataset X, the student $\mathcal{S}$ is trained by not only the teacher $\mathcal{T}$ to minimize the perturbed distillation loss $\tilde{L_{DS}}$, but also the discriminator $\mathcal{D}$ to minimize the adversarial loss L_{AD}. In the following, We first discuss the knowledge transfer path of KD and

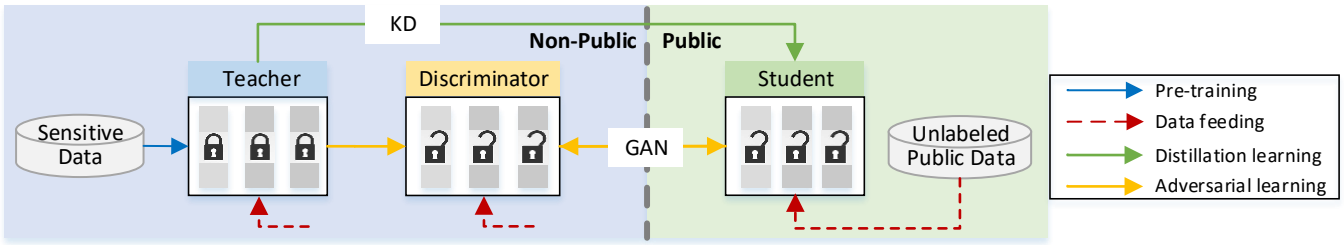


Figure 1: Overview of the proposed framework and its data flow.

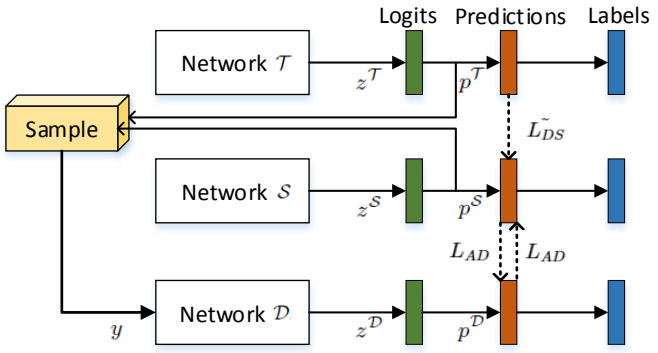


Figure 2: Private knowledge transfer strategy: the student network is trained with adversarial and distillation losses; the discriminator network is trained to distinguish the sampled probability distributions of teacher and student.

its training loss. Then, we introduce the proposed privacy protection mechanism. After that we describe the knowledge transfer path of GAN and the adversarial learning loss. Finally, we present the joint learning procedure for the framework.

4.1 Knowledge Transfer Path with KD

4.1.1 Knowledge distillation

Given a sample set of N samples $X = \{x_i\}_{i=1}^N$ from M classes, we denote the corresponding label set as $Y = \{y_i\}_{i=1}^N$ with $y_i = \{1, 2, ..., M\}$. The categorical distributions of the two networks $\mathcal{T}$ and $\mathcal{S}$ are probability values of M classes, denoted by $p^{\mathcal{T}}$ and $p^{\mathcal{S}}$, respectively. For a sample x_i, the two probabilities are:

$$p^{\mathcal{T}}(x_i) = softmax(z_i^{\mathcal{T}}) \quad \text{and} \quad p^{\mathcal{S}}(x_i) = softmax(z_i^{\mathcal{S}}) \tag{2}$$

where the logits $z^{\mathcal{T}}$ and $z^{\mathcal{S}}$ are the outputs of the last fully connected layer of the two networks $\mathcal{T}$ and $\mathcal{S}$, respectively.

Kullback Leibler (KL) Divergence is a measure of how one probability distribution differs from another. Here we adopt KL divergence as the distillation loss to distill knowledge from network T to network S. Then we have the following loss function:

$$L_{DS} = \sum_{i=1}^{N} p_{\boldsymbol{\tau}}^{\mathcal{T}}(x_i) \log\left(\frac{p_{\boldsymbol{\tau}}^{\mathcal{S}}(x_i)}{p_{\boldsymbol{\tau}}^{\mathcal{T}}(x_i)}\right) \tag{3}$$

where the probability distribution $p_{\boldsymbol{\tau}}(x_i) = softmax(z_i/\boldsymbol{\tau})$ is the softmax temperature function with $\boldsymbol{\tau}$ as the temperature parameter. The temperature parameter $\boldsymbol{\tau} > 0$ controls how much we want to soften or smooth the class probability predictions from $p^{\mathcal{T}}$. A higher temperature indicates a softer probability distribution generated by the teacher network with privileged knowledge of the differences among the classes.

4.1.2 Differential privacy protection

The proposed privacy protection is built upon a general approach in [10], which can allocate privacy budget to each step and compute the total privacy cost over iterations.

Given a probability sample q, clipping threshold C, noise multiplier m, in [10], the standard procedure of adding Gaussian noise ($i.e.$, applying DP) to a vector to be protected is:

1. Select a subset of records (or samples) R_i, $i \subseteq [1, ..., N]$, with probability q to choose each record. The result of each query for the record is a vector $v^i \in R^D$;
2. Clip each v^i by threshold C and L_2 norm: $\pi_C(v^i) = v^i \cdot \min(1, C/||v^i||_2)$;
3. The sum with noise (or DP) added is then computed by: $\tilde{v} = \frac{1}{qN}(\sum \pi_C(v^i) + \mathcal{N}(0; \sigma^2 I))$.

where $\sigma = m \cdot C$ and can be intuitively understood as the scale of the injected noise.

Unlike the general approach in [10] that applys DP to the gradients during training, the primary privacy concerns of the proposed strategy (as briefly discussed in Sec. 3) is only associated with the private teacher network, in particular, the distillation loss to be queried. Thus, in the proposed framework, we do not need to protect the gradient vector as [10]. Instead, we can choose a batch of samples, then protect a vector of batch distillation loss that is queried from the teacher. We can use the following to compute the batch distillation loss:

$$v^i = p_{\boldsymbol{\tau}}^{\mathcal{T}}(x_i) \log\left(\frac{p_{\boldsymbol{\tau}}^{\mathcal{S}}(x_i)}{p_{\boldsymbol{\tau}}^{\mathcal{T}}(x_i)}\right) \tag{4}$$

To protect the vector, we can then inject Gaussian noise with mean of 0 and standard deviation of $m \cdot C$ to the norm-bounded batch distillation loss. As a result, the batch of differentially-private distillation loss can be defined as:

$$\tilde{L_{DS}} = \frac{1}{qN}\left(L_{DS} / \max\left(1, \frac{||L_{DS}||_2}{C}\right) + \mathcal{N}(0; \sigma^2 I)\right) \tag{5}$$

With such a differentially-private loss $\tilde{L_{DS}}$ for each batch of samples, we can bound the privacy cost of each query to the teacher. In addition, it is necessary to keep track of the privacy cost during the entire training procedure and then derive the final privacy budget for a desired model utility. Such a relationship acts as the theoretical basis for our trade-off between utility and privacy. We employ the concept of Rényi divergence and Rényi differential privacy (RDP) accountant [12, 13] as the measure. Per the definitions of DP and RDP, the privacy loss can be considered as a random variable dependent on the injected random noise. RDP accountant then generalizes the pure ϵ-differential privacy and achieves a more compact composition theorem than the standard (ϵ, δ)-DP [12, 13]. In particular, RDP ensures that a randomized mechanism can be bounded by a smaller

ϵ through Rényi divergence of two adjacent inputs, and then tracks privacy bounds on the moments of the privacy loss. In the proposed DP mechanism, RDP is applied to track the bound on the batch distillation loss with sampled Gaussian noise.

4.2 Knowledge Transfer Path with GAN

With the pre-trained teacher network $\mathcal{T}$, the student network $\mathcal{S}$ can be trained adversarially against the discriminator network $\mathcal{D}$. The student and the discriminator play a min-max game. While the network $\mathcal{S}$ attempts to generate a probability $p^{\mathcal{S}}$ mimicking the distribution of the teacher network $p^{\mathcal{T}}$, the discriminator model tries to distinguish the 'true' label predicted by $\mathcal{T}$ from the pseudo label by $\mathcal{S}$. We then can define the objective function L_{AD} for the min-max game:

$$\min_{s}\max_{d} L_{AD} = E_{y\sim p^{\mathcal{T}}}[\log p^{\mathcal{D}}(y) + E_{y\sim p^{\mathcal{S}}}[\log\left(1 - p^{\mathcal{D}}(y)\right] \quad (6)$$

where $y \sim p^{\mathcal{T}}$ and $y \sim p^{\mathcal{S}}$ are the continuous samples generated from the discrete probability distributions of $p^{\mathcal{T}}$ and $p^{\mathcal{S}}$, respectively; $p^{\mathcal{D}}(y)$ is the probability generated by the discriminator network for a label y.

In the proposed framework, the network $\mathcal{D}$ gets updated by maximizing the objective function L_{AD} in Eq. (6), while $\mathcal{S}$ attempts minimizing L_{AD}, thereby making $\mathcal{D}$ unable to differentiate whether a given label is predicted by S or not. Such a min-max game updates $\mathcal{S}$ and $\mathcal{D}$ alternatively until the equilibrium is reached, $i.e.$, $\mathcal{S}$ learns the distribution of $\mathcal{T}$ given the discrimination of $\mathcal{D}$. As shown in Figure 2, the network D cannot directly take the discrete probabilities from $\mathcal{T}$ and $\mathcal{S}$, which may inevitably result in high variances in the gradients. To reduce the variances for $\mathcal{D}$, we use the Gumbel-Max trick [9] to re-parameterize the generation of the discrete samples to a continuous space. Then, we can conduct sampling approximation to obtain the continuous samples y and formulate L_{AD} with lower-variance gradients.

4.3 Joint Learning Procedure

Based on the two knowledge transfer paths and the proposed privacy protection mechanism, we incorporate the differentially private distillation loss in Eq. (5) and adversarial loss in Eq. (6) into the final objective function for the target student as below:

$$L = \alpha\tilde{L_{DS}} + (1-\alpha)L_{AD} \quad (7)$$

where α is a distillation weight set between 0 and 1. We achieve the joint utility and privacy optimization with the proposed privacy protection mechanism that transfers knowledge from a protected private teacher to a public student. The overall learning procedure is summarized in Algorithm 1.

Our learning strategy has its unique advantages: when combined with the adversarial learning, the joint optimization of distillation loss and adversarial loss can greatly save training epochs, which results in a stronger privacy guarantee. In addition, with RDP accountant, by searching the optimal values for the hyper-parameters ($i.e.$, q, C, σ) and the distillation weight α, we can ensure an optimal model utility with a tight privacy bound.

5 EXPERIMENTAL RESULTS

The proposed framework is applicable to a wide range of multi-label learning tasks, where the students can learn from the teachers owning sensitive private data. Note that the privacy budget of a complete training procedure greatly depends on the noise injected to each training step and the number of training epochs. To prove the generality, we here employ three widely-adopted datasets, MNIST, SVNH, CIFAR, to evaluate the proposed framework. We compare our results with the state-of-art existing works in [15, 20] using the reported data of [15] on MNIST and SVHN, and the reported data of [20] on CIFAR-10, respectively.

Algorithm 1: Privacy-Preserving Learning Procedure.

Input: a pre-trained network $\mathcal{T}$, public training samples N, training epochs T, training epochs for the discriminator $T_{\mathcal{D}}$, training epochs for the student $T_{\mathcal{S}}$, batch size B, clipping threshold C, the noise multiplier m.

Output: a student network $\mathcal{S}$

for $t = 0$ *to* $T-1$ **do**
  for $i = 0$ *to* $(T_D - 1) * (N/B)$ **do**
    Sample a batch x of size B;
    Sample y from discrete probabilities $p^{\mathcal{T}}$ and $p^{\mathcal{S}}$;
    Compute adversarial loss L_{AD};
    Update $\mathcal{D}$ by ascending along its gradients of L_{AD} ;
  for $j = 0$ *to* $(T_S - 1) * (N/B)$ **do**
    Sample a batch x of size B;
    Compute distillation loss L_{DS};
    Apply differential privacy mechanism:
    $\tilde{L_{DS}} \leftarrow L_{DS}/\max(1, \frac{||L_{DS}||_2}{C}) + N(0; \sigma^2 I)$;
    Sample y from discrete probabilities $p^{\mathcal{T}}$ and $p^{\mathcal{S}}$;
    Compute adversarial loss L_{AD};
    Compute weighted sum L;
    Update $\mathcal{S}$ by desending along its gradients of L;

5.1 Experiment Setup

Here we briefly describe our experiment setup. We implement our learning strategy based on Tensorflow. The experiments are based on MNIST, SVHN and CIFAR-10 classification tasks. We first pre-train a teacher model on the entire dataset (with separate training and testing), and treat it as the private teacher. Then we randomly select the public data from the training dataset and assume that those data are unlabelled, which is used to train the student model through the proposed learning strategy in Sec. 4.

Three datasets of MNIST, SVNH and CIFAR-10 are used in our experiments with the following details:

MNIST. The MNIST dataset [7] has 60000 grayscale images (50000 for training and 10000 for testing) with 10 different label classes. Teacher, student and discriminator are implemented using a LeNet, an MLP and a LeNet. When trained on the entire dataset, the teacher model has a 99.40% test accuracy. We vary the number of unlabelled training instances in [100, 10000] to train the student.

SVHN. The SVHN dataset [14] consists of 32×32 colored digit images, each digit representing one class. The training and testing sets contain 604388 and 26032 images, respectively. Teacher, student and discriminator are implemented using a ResNet, a LeNet and a LeNet. The teacher has a 95.30% test accuracy after training. The number of unlabelled training instances to train the student is varied in [500, 50000].

CIFAR. The CIFAR-10 dataset [6] contains colored natural images with a size of 32×32, with 10 classes. The training and testing

sets contain 50000 and 10000 images, respectively. The three networks and training setup are the same as the SVHN dataset. The teacher can reach a 89.4% accuracy after training.

5.2 Results and Analyses

The typical application of knowledge distillation is to transfer from a powerful and large network or an ensemble of networks to a small network (also known as deep model compression), which can reduce the network complexity and capacity to improve the deployability of the deep models [20, 21, 22]. In this paper, instead of focusing on the model compression performance, **the goal is to achieve a good student accuracy under a tight privacy bound,** $i.e.$**, optimal trade-off between utility and privacy.** We conduct experiments with different number of unlabelled data to explore how utility and privacy budget vary against the number of training instances and training epochs. This is aligned with the aforementioned motivation of this paper and the practical demands from mobile/edge scenarios.

In the following, we thoroughly study the impacts from training size, epochs, and privacy bounds on the framework performance. For the other hyper-parameters in Algorithm 1 ($e.g.$, batch size, distillation weight, clipping threshold, noise multiplier, $etc.$), the optimal values within the range are searched and pre-determined. It is noted that, even with such practical constraints of tight privacy bound, fewer training epochs and limited public data, the proposed framework still can exhibit higher model utility than prior works, with different network typologies employed to teachers and students ($e.g.$, a MLP student and a LeNet teacher for MNIST). This indicates the overall superiority of the framework and its applicability to real-world mobile/edge scenarios with limited resources.

Training speed. We first investigate the training speed of the framework on the aforementioned three datasets. The learning curves of both teachers and students are plotted in Figure 3, where the students are trained with a subset of 10000 public unlabelled training instances. It is intuitively understandable that, in knowledge distillation, the student model accuracy learned from the teachers is sub-optimal to the teacher. Thus, we observe in the figure that the student models for MNIST, SVHN, CIFAR can quickly reduce the accuracy loss to 0.89% after 20 training epochs, 2.34% after 20 training epochs, and 13.69% after 30 training epochs, respectively. As discussed in Sec. 4, Figure 3 also validates that our student models can reach the convergence with a small number of training epochs (20-30). Thus, the proposed learning strategy that combines KD with GAN greatly **speeds up the training procedure by reducing the number of training epochs for convergence**. As a result, the fewer accesses to the private teacher induces a smaller privacy bound and cost to facilitate a meaningful utility .

Training size. The proposed strategy combines the advantages of both KD and GAN: 1) KD requires a small number of training instances to distill knowledge; 2) GAN enforces the student to learn the true distribution in the min-max game. We compare the proposed joint optimization with the KD-only optimization to show the efficiency of the proposal, the results of which are summarized in Table 1. We vary the size of training sets on MNIST, SVHN and CIFAR-10 and compute the average accuracy over 20 runs. It is observed that the joint optimization of KD and GAN consistently outperforms the KD-only optimization, especially when only a small number of training images are available. For example, there are a 22.99% accuracy improvement for SVHN and a 26.74% accuracy improvement for CIFAR-10 when with 500 training instances. Thus, the joint optimization of KD and GAN **not only reduces the required number of training instances but no longer need to purposely select the training instances** to achieve equally effective student model. This is especially beneficial to healthcare and finance applications where local agents (students) typically have very few data for training.

Utility $v.s.$ **privacy.** To study the trade-off between utility ($i.e.$, student model accuracy or accuracy loss $w.r.t.$ the teacher) and privacy ($i.e.$, differential privacy bound ϵ), we track the values of ϵ for a given failure probability δ in the proposed differential privacy mechanism through the RDP accountant. We set the experiments with training size of 10000, batch size B of 50, noise multiplier m of 1.1. In Figure 4, we report the values of the ϵ-differential privacy bounds (δ is 10^{-5} on MNIST, 10^{-6} on SVHN and CIFAR-10) for student training and testing. It is found that with a smaller ϵ guarantee, our MNIST student can still achieve a high accuracy ($i.e.$ (1.93, 10^{-5})-differential privacy for a 98.52% accuracy). For SVHN and CIFAR-10, student accuracies generally increase with growing privacy bound but quickly converge after certain privacy bound thresholds ($i.e.$, SVHN student's accuracy stays almost the same for privacy bound $\epsilon > 3.5$, and CIFAR-10 student's accuracy for a bound $\epsilon > 8.5$). Thus, the proposed framework is capable to **effectively make trade-off between utility and privacy, allowing the student to achieve an optimal accuracy within a small privacy bound**.

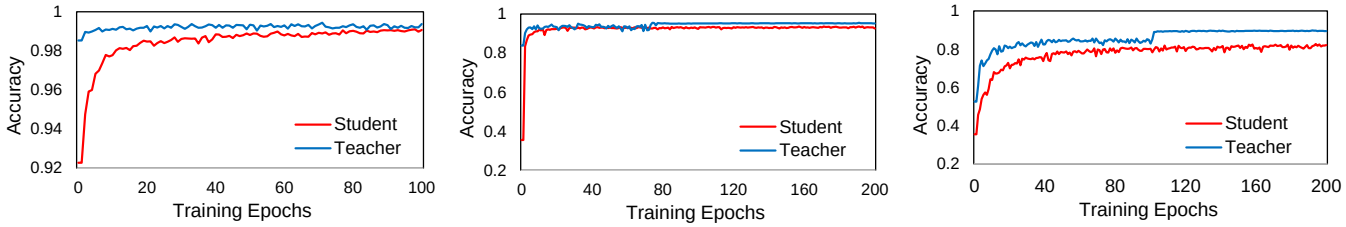
5.3 Comparisons and Discussions

We compare the proposed framework with two state-of-art related works, PATE [15] and RONA [20]. As is discussed in the previous sections, the major differences among the three methods are: for PATE, the student receives knowledge from an ensemble of teachers, whereas for RONA, the student is trained with cross entropy loss supervised from labelled data in addition of distillation loss. As in Sec. 5.2, the proposed framework has already been proved to achieve fast training speed with fewer training instances (that can be unlabelled), exhibiting superiority for privacy-concerned applications in mobile/edge scenarios. Thus, the focus of comparison here is placed upon the utility and privacy trade-off.

Table 2 compares the performance of the three methods, which reports the student accuracies and the accuracy losses $w.r.t.$ teachers for a given privacy bound. For fair comparison, the reported data of PATE and RONA in Table 2 are directly taken from the papers [15, 20] to prevent any re-implementation issues. Note that RONA requires all the public data are labelled [20], which is different from the setup of PATE and the proposed framework. However, since PATE does not have the results for CIFAR-10, we compare with RONA on CIFAR-10 to demonstrate the scalability of the proposed framework (even under unfavorable conditions). It is observed that the proposed framework can offer better trade-offs between privacy (privacy bound ϵ) and utility (student model accuracy) than the other two methods for all the datasets. Moreover, with the relaxation of privacy bound ϵ, the proposed framework can achieve more significant accuracy improvement than the other works. For examples, when ϵ is increased from 2 to 8 for MNIST, the accuracy loss of our student is reduced from 0.89% to 0.20%, while the change for PATE is very small, from 1.21% to 1.11%, indicating an 82% and 26% *accuracy loss improvements* (defined as the difference between the accuracy losses for two methods over the loss for the reference method) for the two privacy bounds.

Such performance difference is due to the fact that the the accuracy improvement of PATE requires a massive number of queries, while its privacy budget sharply grows with increasing queries [15].

Figure 3: Training curves (blue) of teachers on the private dataset, and training curves (red) of students on a public subset with 10000 training instances for the three datasets: (a) MNIST; (b) SVHN; (c) CIFAR-10.



(a) Training curves for MNIST (b) Training curves for SVHN. (c) Training curves for CIFAR-10.

Table 1: Comparison on average accuracy over 20 runs between the proposed joint learning strategy (combining KD and GAN) and the KD-only optimization for different number of training instances (denoted as n) for students.

Method	MNIST			SVHN			CIFAR		
	n=100	n=1000	n=10000	n=500	n=5000	n=50000	n=500	n=5000	n=50000
KD-only (%)	66.61	93.57	98.38	33.24	86.28	92.49	19.71	64.59	81.18
Joint (%)	67.99	96.46	98.52	56.23	87.74	96.62	46.45	66.9	80.06

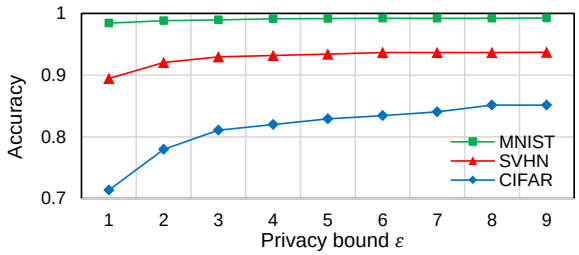


Figure 4: Student model accuracy $v.s.$ privacy bound ϵ for the three datasets.

On the other hand, the proposed mechanism applies Gaussian noise with random sampling to the loss function for DP guarantee and is shown with good scalability, as the privacy cost of single query is quadratically reduced with the sampling rate. Such scalability limitation of PATE can be amplified when with more complex dataset. For example, for SVHN, with an ϵ of 5.04, the student model from PATE has an accuracy of 82.70% $w.r.t.$ a 92.80% teacher accuracy, resulting in significant accuracy loss. Unlike PATE, with a similar ϵ of 5.02, our framework can achieve 93.12% student model accuracy for a 95.30% teacher accuracy ($i.e.$, 79% accuracy loss improvement from PATE), making it feasible to achieve a desired accuracy and privacy trade-off.

For the comparison with RONA, with ϵ bounds of 4.21 and 8.81, our student can achieve accuracies of 81.76% and 84.79%, which is slightly better than the performance of RONA[20] ($i.e.$, 5% accuracy loss improvement from RONA). However, as is discussed earlier, RONA relies on both transfer and self learning where all the public samples must be labelled. Such scenarios can be limited in practice. Unlike that, our framework can achieve good performance even with training data unlabelled, which is a more challenging but practical scenario, indicating a broader applicability.

Note that the proposed framework can be used as a baseline backbone for private knowledge transfer even with limited quality public data, which is well aligned with the goal of this paper to develop the transfer framework instead of extensive optimizations. In other words, many other optimization techniques, such as well-designed sample selection for querying, carefully-analyzed sensitivity, can be seamlessly integrated into this framework to further improve model utility or trade-off between utility and privacy. For example, the selective aggregation mechanism in improved PATE [16] can reduce the added noise and achieve an accuracy loss of 0.71% for a (1.97, 10^{-5}) bound on MNIST. Such optimization techniques are orthogonal to the proposed privacy knowledge transfer framework and hence can be always employed at the cost of complexity, which is not the focus of the paper, but can be explored in the future.

Table 2: Utility and privacy trade-off comparisons among the proposed framework (Proposed), PATE [15] and RONA [20]

.

Dataset	Framework	Privacy Bound ϵ	Accuracy		Accuracy loss
			Student	Teacher	
MNIST	PATE	2.04 8.03	98.0% 98.1%	99.20%	1.21% 1.11%
	Proposed	1.93 8.00	98.52% 99.20%	99.40%	0.89% 0.20%
SVHN	PATE	5.04 8.19	82.70% 90.70%	92.80%	10.88% 2.26%
	Proposed	5.02 8.18	93.12% 93.71%	95.30%	2.29% 1.68%
CIFAR	RONA	4.20 8.87	78.6 % 81.69%	86.35%	8.98% 5.40%
	Proposed	4.21 8.81	81.76% 84.79%	89.40%	8.54% 5.16%

6 CONCLUSION

This paper presented a three-player (teacher-student-discriminator) framework that transfers private knowledge and improves privacy and utility trade-off. The proposed framework combines KD from a teacher with GAN involving a student and a discriminator. A key insight in the proposed learning strategy is that the combination of KD and GAN provides additional quality supervision in addition to the distilled knowledge. Then, with a differential privacy mechanism, the proposed framework can establish a precise privacy guarantee of the training procedure while the training convergence can be quickly achieved even with limited training epochs and unlabelled training instances. Experimental results show that the proposed framework can act as the baseline framework for private knowledge transfer and achieve excellent utility and privacy trade-off on the datasets of MNIST, SVHN and CIFAR-10 for multi-label classification tasks.

7 ACKNOWLEDGEMENTS

This work was supported in part by the National Key R&D Program of China (Grant No. 2018YFE0126300) and the National Science Foundation of China (Grant No. 61974133).

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, 'Deep learning with differential privacy', in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318. ACM, (2016).

[2] Cynthia Dwork, 'Differential privacy', *Encyclopedia of Cryptography and Security*, 338–340, (2011).

[3] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith, 'Calibrating noise to sensitivity in private data analysis', in *Theory of cryptography conference*, pp. 265–284. Springer, (2006).

[4] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart, 'Model inversion attacks that exploit confidence information and basic countermeasures', in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333. ACM, (2015).

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, 'Image-to-image translation with conditional adversarial networks', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, (2017).

[6] Alex Krizhevsky, Geoffrey Hinton, et al., 'Learning multiple layers of features from tiny images', Technical report, Citeseer, (2009).

[7] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al., 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE*, **86**(11), 2278–2324, (1998).

[8] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik, 'Unifying distillation and privileged information', *arXiv preprint arXiv:1511.03643*, (2015).

[9] Chris J Maddison, Daniel Tarlow, and Tom Minka, 'A* sampling', in *Advances in Neural Information Processing Systems*, pp. 3086–3094, (2014).

[10] H Brendan McMahan and Galen Andrew, 'A general approach to adding differential privacy to iterative training procedures', *arXiv preprint arXiv:1812.06210*, (2018).

[11] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang, 'Learning differentially private recurrent language models', *arXiv preprint arXiv:1710.06963*, (2018).

[12] Ilya Mironov, 'Rényi differential privacy', in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, (2017).

[13] Ilya Mironov, Kunal Talwar, and Li Zhang, 'R\'enyi differential privacy of the sampled gaussian mechanism', *arXiv preprint arXiv:1908.10530*, (2019).

[14] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng, 'Reading digits in natural images with unsupervised feature learning', (2011).

[15] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar, 'Semi-supervised knowledge transfer for deep learning from private training data', *arXiv preprint arXiv:1610.05755*, (2016).

[16] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson, 'Scalable private learning with pate', *arXiv preprint arXiv:1802.08908*, (2018).

[17] Reza Shokri and Vitaly Shmatikov, 'Privacy-preserving deep learning', in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321. ACM, (2015).

[18] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov, 'Membership inference attacks against machine learning models', in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, (2017).

[19] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart, 'Stealing machine learning models via prediction apis', in *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pp. 601–618, (2016).

[20] Ji Wang, Weidong Bao, Lichao Sun, Xiaomin Zhu, Bokai Cao, and S Yu Philip, 'Private model compression via knowledge distillation', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1190–1197, (2019).

[21] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi, 'Kdgan: knowledge distillation with generative adversarial networks', in *Advances in Neural Information Processing Systems*, pp. 775–786, (2018).

[22] Zheng Xu, Yen-Chang Hsu, and Jiawei Huang, 'Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks', *arXiv preprint arXiv:1709.00513*, (2017).

[23] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu, 'Seqgan: Sequence generative adversarial nets with policy gradient', in *Thirty-First AAAI Conference on Artificial Intelligence*, (2017).

[24] Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, and Stacey Truex, 'Differentially private model publishing for deep learning', *arXiv preprint arXiv:1904.02200*, (2019).

[25] Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin, 'Adversarial feature matching for text generation', in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4006–4015. JMLR. org, (2017).