IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received March 9, 2019, accepted March 25, 2019, date of publication April 9, 2019, date of current version April 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2909559

Differential Privacy Preservation in Deep Learning: Challenges, Opportunities and Solutions

JINGWEN ZHAO[1], YUNFANG CHEN[1], AND WEI ZHANG[1,2], (Member, IEEE)
[1]School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[2]Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Corresponding author: Wei Zhang (zhangw@njupt.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61672297.

ABSTRACT Nowadays, deep learning has been increasingly applied in real-world scenarios involving the collection and analysis of sensitive data, which often causes privacy leakage. Differential privacy is widely recognized in the majority of traditional scenarios for its rigorous mathematical guarantee. However, it is uncertain to work effectively in the deep learning model. In this paper, we introduce the privacy attacks facing the deep learning model and present them from three aspects: membership inference, training data extraction, and model extracting. Then we recall some basic theory about differential privacy and its extended concepts in deep learning scenarios. Second, in order to analyze the existing works that combine differential privacy and deep learning, we classify them by the layers differential privacy mechanism deployed, such as input layer, hidden layer, and output layer, and discuss their advantages and disadvantages. Finally, we point out several key issues to be solved and provide a broader outlook of this research direction.

INDEX TERMS Deep learning, differential privacy, privacy attacks.

I. INTRODUCTION

In the era of big data, the explosive growth of data volume has accelerated the development of deep learning. Recently, as a state-of-the-art field of machine learning research, deep learning has achieved remarkable success. Relying on its capabilities of multi-level representation and abstraction, deep learning can understand data, such as images, sounds, texts and others [1], and its applications are being expanded into various areas, such as social network analysis [2], internet of things [3], [4], bioinformatics [5], wireless communications [6], [7], medicine and healthcare [8], malware detection [9] and so on.

The issue of privacy protection attracts increasing attention with the accelerated integration of the big data industry and people's daily life. In terms of policy, the General Data Protection Regulation (GDPR) [10] came into force on May 25th, 2018 in all EU member states to harmonize data privacy laws across Europe, which essentially set a new global standard for data protection. The GDPR aims primarily to give individuals more control over their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EU. Moreover, from a technical aspect, privacy protection research has never stopped. Dalenius [11] proposed the concept of private disclosure control, and the k-anonymity algorithm [12] lays a foundation for the anonymous privacy protection algorithm based on equivalence class grouping, followed by l-diversity [13], t-closeness [14], (α, k)-anonymity [15] and so on. These models improve the anonymity protection theory against attackers with different background knowledge. Nevertheless, all of them have some common defects, which need to update design to catch the fast-evolving attacks, and cannot provide strict proof to quantify the privacy protection effect.

The differential privacy (DP) proposed by Dwork *et al.* [16] in 2006 has shown provable privacy guarantees for database record releasing without significant query accuracy loss, even if the adversary possesses all the remaining tuples of the sensitive data. Several attempts have been taken to apply differential privacy into deep learning, in order to guarantee the privacy of samples in training datasets, by combining its strict mathematical proof with flexible composition theorems. Recently, differential privacy has been applied to real-world products. For example, it can be used in user data

The associate editor coordinating the review of this manuscript and approving it for publication was Tomohiko Taniguchi.

collection and analysis such as Spotlight and Notes in input method and search function of Apple's iOS10 [17], [18]; Google's Chrome [19], [20], Samsung's smartphone [21] want to use it to protect user's personal information and data details, while extracting the user's general information needed for machine learning. The differential privacy method to protect user private data has been gradually recognized in the industry as a practical standard for privacy protection.

When individual data (e.g. clinical records, user habits, photos, etc.) are used to train deep learning model, some sensitive features will be 'remembered'. Deep learning algorithms typically use regularization techniques (e.g. l_2 regularization, dropout, etc.) to prevent model over-fitting, which is helpful to protect the privacy of training data. However, it is insufficient for the deep models with excellent learning ability to prevent them from remembering the privacy details. Recently, attacks using implicit memory in machine learning have shown that sensitive training data can be recovered from the model. This type of attack can be performed directly by analyzing internal model parameters, or indirectly by querying in black-box setting repeatedly. In order to improve the security of deep learning model by deploying the differential privacy mechanism, how much to add and where to add noise to the deep learning network requires careful consideration, because any subtle changes will make the prediction very different under the network layer-by-layer abstraction. Furthermore, because differential privacy needs to be irrelevant to background knowledge, it is difficult to control the privacy budget within a reasonable range in reality. For example, there are reports showing that a large privacy budget is actually used in Apple [22]. Therefore, how much protection that differential privacy can provide for deep learning is worthy of further discussion.

The more diverse the application scenarios of deep learning model, the more serious the privacy security issues will be. How to integrate differential privacy mechanism and deep learning effectively becomes a research hotspot. The contribution of our paper can be summarized as the following three aspects:

1) According to the different effects of privacy attack on deep learning, we classifies the privacy threats of deep learning, and clarifies how differential privacy be applied to the specific scenarios.
2) According to the deep learning network structure, we divide schemes into three categories. In addition, we summarize the core idea of every method and its specific implementation, and point out the advantages and disadvantages of them.
3) We provide several possible research directions, which gives a useful reference for researchers to explore the differentially private deep learning mechanism.

The rest of the paper is organized as follows. Section II introduces the privacy threats faced by deep learning model. Section III demonstrates the basic theory of differential privacy and its extensive concepts. Section IV illustrates several concrete schemes according to the position of layer in deep learning model that differential privacy applied. Section V puts forward a few issues of privacy protection for deep learning, and enlightens the future research. Section VI concludes the paper.

II. PRIVACY ISSUES IN DEEP LEARNING

Deep learning as a state-of-art technology in machine learning significantly improves the classification accuracy on highly-structured and large-scale database, because it enables the deployment of end-to-end learning systems where features and classifiers are learned at the same time. Deep learning uses multiple nonlinear layers transformation to perform representation learning, and it has powerful capabilities of data abstraction [23].

Internet giants like Google, Amazon, Microsoft, etc. are offering 'ML-as-a-Service' (MLaaS), a black-box API that provides deep learning service. By using MLaaS, users can get predictive results through uploading dataset and possessing clustering or regression tasks. The machine learning API make prediction by using the features of input samples, which usually contain sensitive information. The reason for these privacy leaks is mainly because of over-fitting, that means the model implicitly memorizes some details about training data [24], and it is also associated with the structure and type of the model itself. The privacy threats in deep learning models can be classified according to training phase and prediction phase.

During training phase, the privacy threats are closely related to deep learning deployment structure. In centralized learning, model takes a lot of advantage of accuracy by collecting a large amount of data for training. However, it also brings high load to centralized server, and once attacks happens all individual data will be at risk, as shown in the Fig. 1(a). Moreover, the privacy of the ML model and data are independent of how the ML model is used, but rather the extent of the adversary's access to the system hosting the model and data, so it can be seen as a traditional access control problem.

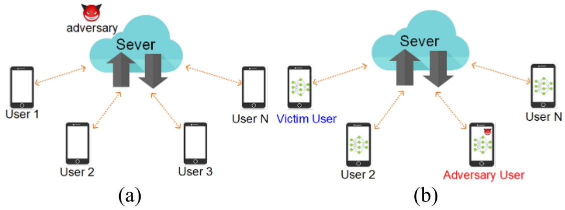

FIGURE 1. Two different ways of model training. (a) Centralized learning. (b) Collaborative learning.

Recently, collaborative learning has been proposed, in which local users and centralized server take part of the training tasks respectively, and only share a subset of the parameters. Unfortunately, in case of malicious participants existing, as shown in Fig. 1(b), they can train the Generative Adversarial Networks (GAN) [25] for information theft. Hitaj *et al.* [27] proposed an attack against the cooperative deep learning, in which the adversary trains a GAN to

TABLE 1. Privacy attacks.

Knowledge of model	Access to model input and output	Access to training data	Privacy threat		
			Membership inference	Training data extraction	Model extraction
White-box	Full	No		[24]	
	Through pipeline only	No			
Black-box	Yes	No	[26]	[29][30]	[31]
	Input only	Yes			[31]
	Through pipeline only	No			

generate prototypical samples with the same distribution of the private target training set. During the training phase, the malicious user is always active, and deceives the victims to release their private information.

During prediction phase, Doshi-Velez and Kim [28] discussed the privacy attacks in three aspects: membership inference, training data extraction and model extracting, as shown in Table 1.

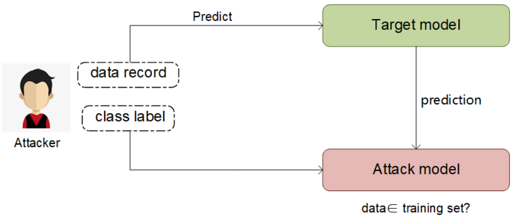

FIGURE 2. Membership inference attack.

A. MEMBERSHIP INFERENCE ATTACK

Membership inference attack is proposed in [26], which is in the black-box setting. The adversary infers whether the record exists in the training dataset or not, giving a machine learning model and a certain sample record. As shown in Fig. 2, the attacker queries the target model with a data record to obtain a prediction on this record, which is a vector of confidence. Then the vector along with the label of the target record is passed to construct the attack model. Because of the different results of the target model processing training sample and unseen sample, the attack model can identify the difference and know whether the record was 'in' or 'out' of the target model's training set.

Training the attack model is first to construct a number of 'shadow models', which are similar to the target model. The author uses supervised training on these shadow models, explicitly teaching them to tell the corresponding output from member or non-member in training dataset (with the label 'in' or 'out'). In addition, they train these shadow models on synthesizes data, that have similar statistical features to training data. Finally, multiple shadow models are cooperatively trained to construct the attack model.

For example, knowing that a machine learning model is used to determine the appropriate drug dose or to discover the genetic basis of a disease [29], it is sufficient to derive the conclusion that if a patient has the disease. The defense principle for membership inference attack and the differential privacy mechanism are conceptually the most similar, and they both aim at the protection for the existence of one sample. Therefore, most of the current differential privacy protections for deep learning models are used to against membership inference attack.

B. TRAINING DATA EXTRACTION

In white-box setting, there is privacy threat of training data extraction. Ateniese *et al.* [24] defined a meta-classifier model that can be trained to extract useful information from the target classifier. The authors successfully complete several attacks against existing ML classifiers: attacks on network traffic classifiers that implementing support vector machines (SVMs), and speech recognition software based on hidden Markov models (HMM), and they prove that the full disclosure of the algorithm details will result in privacy risks.

With the extensive application of MLaaS, privacy attacks in black-box setting will bring significance that is more practical. Fredrikson *et al.* [29] constructed model inversion attacks for deep models, using the output of the model to infer certain features of the training set. Using patient data as training set, they pointed out the association between drug dose and patient gene. However, for the reason of the inherent medical facts between the two elements, some people hold the opinion that this attack does not cause privacy leakage. In [30], the confidence score provided by the facial recognition system API is used to construct an attack model. The user's recognizable image can be obtained by simply accessing the face recognition system and inquiring with the name of the user. The attack effect is shown in Fig. 3.

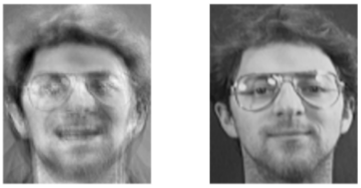

FIGURE 3. An image recovered using a new model inversion attack (left) and a training set image of the victim (right).

However, the model inversion attack only obtains the fuzzy features of the same type of input, that is, just the average of all objects in a given class is generated rather than an explicit data record. If the training set is replaced with a different type of picture, then the result has nearly no effect [26]. As shown in Fig. 4, the result images in top line are airplane, automobile, bird, cat, deer, and the bottom are dog, frog, horse, ship, truck. These images cannot correspond to any specific sample from the training dataset, and even human cannot recognize its category. Therefore, model inversion attack does not give a clear and specific description of the training set, nor can it judge whether a sample is in the training set or not. The attack effect is very limited, that is not suitable for any tasks except face recognition and only available for some particular types of model.

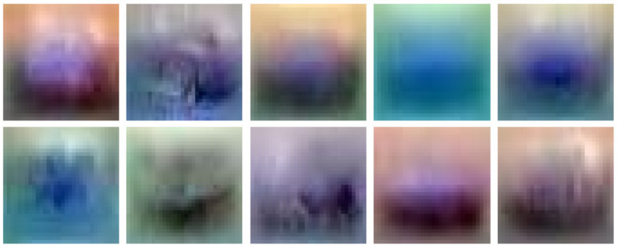

FIGURE 4. Images produced by model inversion on a trained CIFAR-10 model.

C. MODEL EXTRACTION

The model extraction attack is designed to extract the parameters of the model trained on the private data. The attacker aims to duplicate the functionality of the model, whose prediction performance on verification data set is similar to the target model. Because of the close connection between the model parameters and the training set, the privacy of the training set will be further revealed after the leakage of the parameters of the black-box model. A model $\hat{f}$ which is similar with the target model f can be constructed by continuously providing the sample to the black box model and recording the prediction vector [31]. Then depending on solving equations or path-finding algorithm, the decision tree for original data reconstruction can be obtained. As shown in Fig. 5, model extraction attacks also can serve as a 'stepping stone' for other privacy attacks to enhance their effects, such as model inversion.

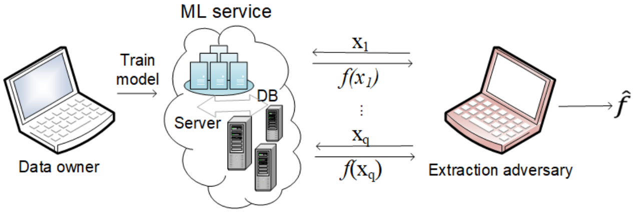

FIGURE 5. Model extraction attack.

III. DIFFERENTIAL PRIVACY PRELIMINARIES

A. DEFINITION

Differential privacy [32] provides a strong privacy guarantees for algorithms processing aggregate database. It is defined in the context of adjacent databases, which differ in a single data record. In deep neural networks, adjacent databases exist in training dataset, which is consist of many image-label pairs [33]. Two of databases are adjacent, if only one image-label pair is present in one database and absent in the other.

The definition of differential privacy is as follows: A randomized mechanism $M : D \rightarrow R$ with domain D and range R satifies (ε, δ)-differential privacy [34], if for any two adjacent inputs $d, d' \in D$ and for any subset of outputs $S \subseteq R$ it holds that

$$Pr[M(d) \in S] \leq e^{\varepsilon} Pr\left[M\left(d'\right) \in S\right] + \delta \tag{1}$$

The trade-off between the accuracy and privacy leakage of the mechanism M is controlled by adjusting the privacy budget parameter ε. A smaller the privacy budget represents a less privacy leakage and a stronger privacy level. The additional variant δ, introduced in [34], allows for the possibility that plain ε-differential privacy is broken with probability δ, which is preferably smaller than $1/|d|$. If δ is 0, the randomized mechanism M gives ε-differential privacy by its strictest definition.

B. COMPOSITION THEOREM

Differential privacy is characteristic of two privacy budget composition theorems: sequential composition [35] and parallel composition [36]. In the application of differentially private deep learning, the sequential composition is widely used.

1) Sequential Composition

Suppose a set of randomized privacy mechanisms $M_i(1 \leq i \leq n)$ sequentially performed on a dataset D and each M_i provides ε_i-DP, they will provide ε-DP, in which $\varepsilon = \sum_{i=1}^{n} \varepsilon_i$.

2) Parallel Composition

Suppose a set of randomized privacy mechanisms $M_i(1 \leq i \leq n)$ and a dataset D divided into several disjointed subsets $\{D_1, D_2, \ldots D_n\}$, the privacy mechanism M_i provides ε_i-DP for every D_i, and they will provide $(max\{\varepsilon_1, \ldots, \varepsilon_n\})$-DP on the entire dataset.

C. SENSITIVITY

In [32], for a query $f : D \rightarrow R$, and neighboring datasets D and D', the sensitivity of f is defined as

$$\Delta f = \max_{D,D'} \left\| f(D) - f\left(D'\right) \right\|_1 \tag{2}$$

Sensitivity considers the maximal difference between the query results on neighboring datasets, which is the change of output of query caused by the single sample in worst case. Sensitivity Δf is only related to the query f and the distribution of the data set, providing a benchmark for the addition of perturbation.

D. PRIVACY LOSS

Privacy loss is a random variable dependent on the random perturbation added to the algorithm and Dwork *et al.* [34]

provided a specific definition of it. For the neighboring databases $d, d' \in D$, a differentially private mechanism $M : D \to R$, auxiliary input aux, and an outcome $o \in R$, the privacy loss at o is

$$\mathrm{c}\left(\mathrm{o}; \mathrm{M}, \mathrm{aux}, \mathrm{d}, \mathrm{d}'\right) \triangleq \log \frac{\Pr\left(\mathrm{M}\left(\mathrm{aux}, \mathrm{d}\right) = \mathrm{o}\right)}{\Pr\left(\mathrm{M}\left(\mathrm{aux}, \mathrm{d}'\right) = \mathrm{o}\right)} \tag{3}$$

Privacy loss is calculated at each step of the algorithm and is accumulated to bound the overall privacy loss of the algorithm, which can be performed by the privacy accountant [36]. The further introduction of privacy accountant will be given in section IV.

E. THE DIFFERENTIAL PRIVACY MECHANISMS

There are two basic mechanisms widely used in deep learning to guarantee differential privacy: the Laplace mechanism [32] and the Exponential mechanism [35].

1) The Laplace Mechanism [32]

For a function $f : D \to R$ over a dataset D, the mechanism M ensures ε-differential privacy, if

$$M(D) = f(D) + Lap\left(\frac{\Delta \mathrm{f}}{\varepsilon}\right) \tag{4}$$

2) The Exponential Mechanism [35]

For non-numeric queries, exponential mechanism randomizes the results, associated with a score function $q(D, \varphi)$. The function q is used to evaluate the quality of the output φ, and different applications lead to various score functions. Δq represents the sensitivity of q. The exponential mechanism M satisfies ε-differential privacy if

$$M(D) = \left(return\ \varphi \propto \exp\left(\frac{\varepsilon \mathrm{q}(\mathrm{D}, \varphi)}{2\Delta q}\right)\right) \tag{5}$$

In addition, noise adding based on a Gaussian distribution [37] is often used in the functional perturbation approach [38]. $N\left(0, S_f^2 \cdot \sigma^2\right)$ is the normal (Gaussian) distribution with mean 0 and standard deviation $\mathrm{S}_f\sigma$. The M mechanism satisfies (ε, δ)-differential privacy if $\delta \geq 4/5exp(-(\sigma\varepsilon)^2/2)$ and $\varepsilon < 1$.

$$M(D) = \Delta f(D) + N\left(0, S_f^2 \cdot \sigma^2\right) \tag{6}$$

F. UTILITY MEASUREMENT OF DIFFERENTIAL PRIVACY

The utility of differential privacy can be measured by the amount of noise and errors. A smaller amount of noise indicates a higher utility. Errors are often measured by accuracy index that depend on utility loss evaluated by the difference between the non-private output and the private output.

IV. DIFFERENTIALLY PRIVATE DEEP LEARNING MECHANISMS

In this section, we focus on several typical differentially private deep learning schemes. Differential privacy guarantees that the output of deep learning model does not show significant statistical differences while the model was trained on adjacent datasets, in which the samples containing individual privacy. The goal of the mechanism is to provide privacy protection for training dataset, preventing privacy leakage in white-box or black-box scenarios. Most of the defense mechanisms are against membership inference attacks mentioned in section II. According to the stage of data processing in deep learning model, we study the mechanisms by dividing the locations of differential privacy deployment into three types: input layer, hidden layer and output layer, as shown in the Fig. 6. The detailed description of Fig. 6 will be given in next parts.

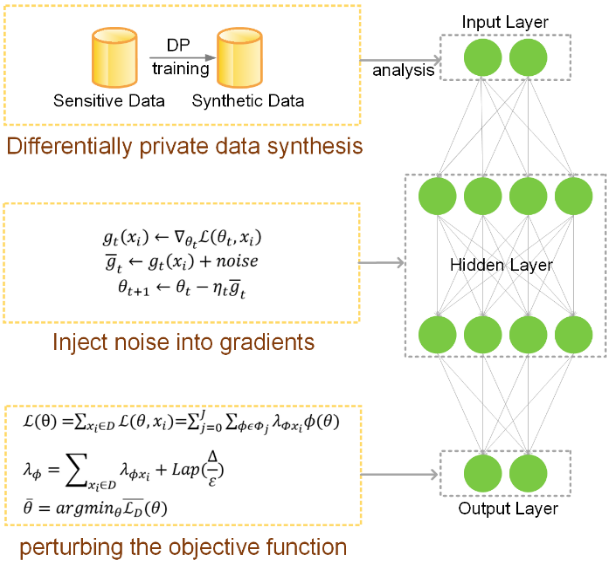
DP
training
Sensitive Data
Synthetic Data
Differentially private data synthesis
analysis
Input Layer
$g_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$
$\overline{g}_t \leftarrow g_t(x_i) + noise$
$\theta_{t+1} \leftarrow \theta_t - \eta_t\overline{g}_t$
Inject noise into gradients
Hidden Layer
$\mathcal{L}(\theta) = \sum_{x_i \in D} \mathcal{L}(\theta, x_i) = \sum_{j=0}^{J} \sum_{\phi \epsilon \Phi_j} \lambda_{\Phi x_i} \phi(\theta)$
$\lambda_\phi = \sum_{x_i \in D} \lambda_{\phi x_i} + Lap(\frac{\Delta}{\varepsilon})$
$\bar{\theta} = argmin_\theta \overline{\mathcal{L}_D}(\theta)$
perturbing the objective function
Output Layer

FIGURE 6. **The three locations of differential privacy deployed in deep learning model.**

A. DIFFERENTIAL PRIVACY DEPLOYED AT THE INPUT LAYER

Differential privacy deployed at the input layer can be seen as a preprocessing of training datasets, which is differently private data synthesis. The data curator firstly generates synthetic data with the same statistical characteristics as the original training datasets under the differential privacy. Then, the synthetic data or generative model is published without privacy leakage, which can be used for various analyses. It guarantees sample-level participant privacy. The solution aims at hiding or changing sensitive features of training datasets, and simultaneously keeping utility for further analysis in deep learning model.

A common approach of generating synthetic data is to add noise to the model directly. Acs *et al.* [39] used a two-stage process that first performs differentially private K-Means clustering [40] for data division, and then produces separately the k models with generative neural model, such as Restricted Boltzmann Machine(RBM) [41] or Variational Auto-encoders (VAE) [42], and trains them on individual group with stochastic gradient descent (SGD) way. The final model consists of the k sub-models. Compared with single generative model, the features details will be learned with higher efficiency in each sub-model, and the noise addition will be more refined, so data utility is improved.

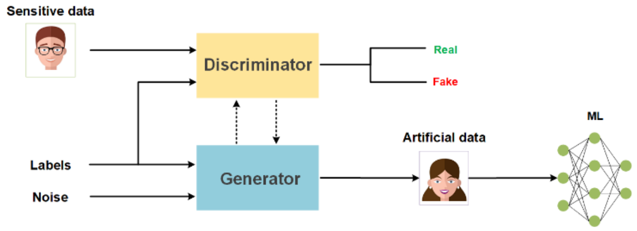

FIGURE 7. **Private artificial data synthesis based on GAN.**

Taking advantage of the GAN, several methods are proposed to generate synthetic data to get better effect [43], [44], as shown in Fig. 7. The results under method of GAN demonstrate that it offers better ability to simulate realistic-looking data that closely matches the distribution of the source data than RBM and VAE. Beaulieu-Jones *et al.* [44] trained the discriminator under differentially private SGD, which generates plausible individuals of clinical datasets. Zhang *et al.* [43] proposed dp-GAN, a general private data publishing framework for rich semantic data (such as image, high- dimensional data) without the requirement of tag information compared to [44].

In order to evaluate the utility of artificial data, we can compare the performances between original data and synthetic data on the same deep learning algorithms. To figure out the potential privacy risks of such technique, Triastcyn and Faltings [45] designed a framework for ex post analysis of generated data. The KL (Kullback–Leibler) divergence estimation and Chebyshev's inequality are used to find a statistical bound on expected privacy loss.

To address the problem of utility degradation of synthetic datasets, relaxation in differential privacy is put forward. A formal privacy guarantee for releasing sensitive datasets is provided [46] by using a criterion called plausible deniability [47]. For any dataset D with $|D| \geq k$, and any record y generated by a probabilistic generative model M such that $y = M(d_1)$ for $d_1 \in D$, we state that y is releasable with (k, γ) *-plausible* deniability, if there exist at least $k-1$ distinct records $d_2, \ldots, d_k \in D\backslash\{d_1\}$ which meet the inequality below

$$\gamma^{-1} \leq \frac{p_r\left\{y = M\left(d_i\right)\right\}}{p_r\left\{y = M\left(d_j\right)\right\}} \leq \gamma, \quad \forall i,\ j \in \{1, 2, \ldots, k\} \tag{7}$$

This mechanism results in input indistinguishability that means by observing the output set (i.e., synthetics) an adversary cannot make sure whether a particular data record was in the input set (i.e., real data). The degree of this indistinguishability is a parameter, and the process can also satisfy differential privacy if we randomize the indistinguishability parameter. The larger privacy parameter k is, the closer to 1 privacy parameter γ is, and the larger the indistinguishability set is for the input data record. Instead of designing the mechanism directly, it achieves differential private by the idea of testing privacy, which rejects ''bad'' samples towards achieving plausible deniability. That is, a synthetic record provides plausible deniability if there exists a set of real data records that could have generated the same synthetic data.

These mechanisms try to generate synthetic datasets with similar statistical properties to the input data and attempt to cover up the key sensitive information, which help to show good illegibility from the original datasets. However, it is a great challenge to operate on large dataset.

B. DIFFERENTIAL PRIVACY DEPLOYED AT THE HIDDEN LAYER

By adding noise to the gradient in the hidden layer to achieve protection for training datasets, it is the most intuitive method to apply differential privacy to deep learning models. It is also the earliest attempt to deploy different privacy into deep from learning model, and its purpose is to prevent the adversary grasping the accurate personal information of the training data by output. There are a number of innovative improvements in these schemes, such as more accurately noise addition and tighter measurement of privacy loss, which bring some significance for model optimization.

Noise adding is performed around the gradient descent process, which can be called the differentially private Stochastic Gradient Descent (dpSGD) algorithm that mainly includes two steps: sanitizer and privacy accountant. Firstly, the step of sanitizer is to limits the sensitivity of each sample by clipping the gradient of the sample, and then to add noise to the gradient in batches before uploading the parameters. The step of privacy accountant is to track the privacy consumption of the entire training.

In the centralized privacy protection model, the confidential problem of the curator is not considered. The curator collects and processes data uniformly under privacy protection mechanism, followed by data release or further analysis. However, when there are attacks on center server or existing of dishonest curator, the sensitive information will be at high risk of leakage. In the local differential privacy (LDP) [20] which bases on distributed model, the users have more control on their own data. Before being contributed to curator, individual data is disturbed locally that provides protection without relying on third party. After that, the sharing models are related to specific algorithms, such as data mining or machine learning algorithms.

Shokri and Shmatikov [48] designed distributed system of deep neural network under differential privacy based on Selective Stochastic Gradient Descent (Selective SGD or SSGD), which provides a new way for end-to-end applications of deep learning on mobile devices. Participants learn neural-network models on their own input data, and they can benefit from other participants who are concurrently learning similar models. Each participant takes turns to upload and download a percentage of the most recent gradients to avoid getting stuck into local minima. The system applies differential privacy to parameter updates by using

TABLE 2. **Privacy budget bound under different composition methods.**

Method	Naive composition	Strong composition	Moments accountant
Privacy budget bound	$(O(qT\varepsilon), qT\delta) - DP$	$(O(q\varepsilon\sqrt{Tlog1/\delta}), qT\delta) - DP$	$(O(q\varepsilon\sqrt{T}), \delta) - DP$

the sparse vector technique, thus it mitigates privacy loss related to both parameter selection and shared parameter values.

The critical problem in such a scheme is that the addition of noise has more influence on model utility along with the increment of training iterations [49]. Therefore, the addition of noise requires accurate weight to keep the balance between privacy and accuracy. The composition theorem illustrates how the privacy protection of a differentially private mechanism degrades under composition of interactive queries. The composition theorem can provide an overall privacy guarantee on the union of all the interactive queries. However, the most of existing composition theorems for DP, including the strong composition theorem [50], just provide a loose bound on privacy spending. As a result, they exhaust a moderate privacy budget very quickly allowing only a few iterations over the training dataset, thereby resulting in the poor utility of deep models.

Abadi *et al.* [33] proposed a state-of-the-art privacy accounting method called moments accountant based on composition theorem, which provides a more tighter bound for privacy loss and can track cumulative privacy loss by implementing independent differential privacy mechanisms. The concept of moments accountant is based on R'enyi differential privacy [53], which allows combining the intuitive and appealing concept of a privacy budget with application of advanced composition theorems. It is a useful analytical tool, compactly and accurately representing guarantees on the tails of the privacy loss compared with the exiting methods.

Specifically, moments accountant adds noise in batches of training data and sets each batch as a lot which size is L. Then each step should satisfies $(O(qT\varepsilon), qT\delta)$-DP, where $q = L/N$ is the sampling rate of a lot. For T rounds iterations, the privacy budget bound of moments accountant under different composition theorems is different from that of naive composition and strong composition, as shown in Table 2.

Obviously, moments accounting get a better result than the previous methods. In particular, it can be proved that the mechanism M satisfies (ε, δ)-DP if $\delta = \min_{\lambda} exp(\alpha_M(\lambda) - \lambda\varepsilon)$ by defining a Moment Generating Function, α_M.

1) Moment Generating Function

For a given mechanism M, we define the λ^{th} moment $\alpha_M(\lambda; aux, d, d')$ as the log of the moment generating function evaluated at the value:

$$\alpha_M\left(\lambda; aux, d, d'\right) = \log \mathbb{E}_{o\sim M(aux,d)}\left[\exp\left(\lambda c\left(o; M, aux, d, d'\right)\right)\right] \quad (8)$$

In the above approaches, there exists the limitation on training epochs due to the small total privacy budget, which leads the limited application on shallow model and small datasets. For large datasets and complex deep learning networks, adding noise to the gradient is tantamount to have a 'heart surgery'. Therefore, researchers should make efforts to find out a privacy protection method that is independent of the number of training epochs.

C. DIFFERENTIAL PRIVACY DEPLOYED AT THE OUTPUT LAYER

In output layer, the model will provide a prediction value $\hat{Y} = f(X)$ on given input in each epoch. The loss function $L(Y, f(X))$ is usually used to evaluate the gap between the predicted value $\hat{Y} = f(X)$ and the true value Y in deep learning, which is a nonnegative real function. The smaller loss function indicates a better prediction performance of the model on training datasets. Therefore, according to the influence of the output result on the loss function, the deployment of differential privacy on the deep learning algorithm can be considered as an optimization problem.

The Functional Mechanism (FM) [54], as an extension of the Laplace mechanism, performs differential privacy by perturbing the objective function of the optimization problem, rather than its results. The direct perturbation to the results has several restrictions, and it only performs on standard types of regression analysis. Deep learning models, such as Deep auto-Encoder (AE) [55], Deep Belief Network (DBN) [56], use different objective functions (e.g., cross-entropy error, energy-based functions) and algorithms (e.g., contrastive divergence, CD) [57]. Specifically, the FM mechanism injects Laplacian noise into the coefficients according to the polynomial approximation of the objective function $f_D(\omega)$, and then releases the model parameter $\bar{\omega}$ that minimizes the objective function $\bar{f_D}(\omega)$.

Phan *et al.* [51] proposed a novel ϵ differential Private Auto-encoder(PA) through analyzing and perturbing the cross-entropy error functions of the data reconstruction and softmax layer. In particular, they approximates the polynomial forms of cross-entropy error functions by using Taylor Expansion [58], and then injects noise into these polynomial forms so that the ϵ-differential privacy is satisfied in the training phases. Phan *et al.* [52] applied Chebyshev expansion [59] to derive the approximate polynomial representation of objective functions in convolutional deep belief network (CDBN), and adds noise to these polynomials. The private convolutional deep belief network (pCDBN) is used to implement human behavior prediction and binary classification tasks. These schemes bases on FM are compared in Table 3.

To improve model performance after adding noise, Phan *et al.* [60] proposed a method of adaptive noise addition. They perturb affine transformations of neurons and loss

TABLE 3. Comparison of differential privacy schemes based on function mechanism.

Solution	Basic model	Objective function	Expansion method
dPA [51]	Deep Auto-Encoders	Cross-entropy error function	Taylor Expansion
pCDBN [52]	Convolutional Deep Belief Network	Energy function	Chebyshev Expansion

functions used in deep neural networks and adaptively inject noise into features based on the contribution of each output to the results. This mechanism intentionally adds ‘more noise’ into features which are ‘less relevant’ to the model output, and vice-versa. By adding noise on the feature, affine transformation layer, and loss function, the overall scheme satisfies the differential privacy. Moreover, it ensures the utility of the model to a certain extent with making privacy budget independent of the number of training epochs, so the model can be applied to a variety of deep neural networks.

For this kind of deep private learning mechanisms, the fundamental theory is ERM (Empirical Risk Minimization) [61], [62] and PAC (Probably Approximately Correct) learning theory [63]. ERM helps to select the best learning model by transforming the learning process into a convex minimization problem. PAC learning estimates the relationship between the number of learning samples and the accuracy of the model. However, there are some limitations in deep private learning: ERM requires that the objective function should be convex and conform to the L-Lipschitz condition; PAC learning can only be applied while the algorithm is PAC learnable, which hinders the practical development of the private learning. Though the theoretical study is active, it is still not suitable for real-world applications. Different deep learning models have their own unique deployment methods, which cannot be easily migrated to other deep learning models. The key problem to be solved in differentially private deep learning lies in the applicability in reality. Thus, it is an urgent need to develop a universal privacy protection framework with more generalization.

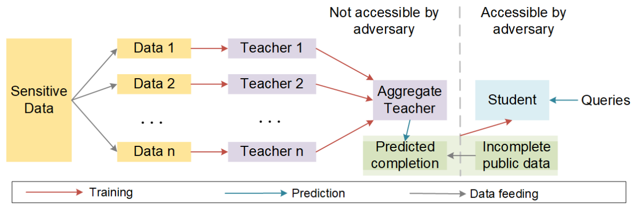

FIGURE 8. The framework of PATE mechanism.

In addition to the above schemes, based on the idea of knowledge aggregation and migration, Papernot *et al.* [64] demonstrated a generally applicable approach for protecting the privacy of training data, the Private Aggregation of Teacher Ensembles (PATE) framework, as shown in Fig. 8. The approach takes advantage of semi-supervised knowledge migration to guarantee the training data privacy in deep learning, and it is applicable to any model, including non-convex deep learning models. This method has been extended in [65] to achieve a more advanced privacy (single digit DP bound) accuracy in the image classification task. The approach trains a large number of ’teacher’ models by assembling disjoint datasets from the sensitive training set to form a ‘teacher ensemble’ that is not published. By adding noise to each teacher’s prediction results, a unified predictive output is formed, which is used to train the ’student’ model. It makes the training on students just relying on the teachers’ predictions without involving their internal parameters.

There are following advantages in PATE. When there are some problems in one ‘teacher’, it does not have a severe impact on the ‘student’; The attacker cannot get the internal structure of the ‘teacher’ through the ‘student’ model; Once the ‘student’ training is completed, the ‘teacher ensemble’ can be removed, as well as the privacy it carries.

The key assumption of this model is that the student model can access unmarked, non-sensitive public data, which statistical characteristics are similarly to the training data teachers used. Nevertheless, we need notice that it is usually difficult to obtain this kind of data in medicine and other realistic fields.

V. FUTURE RESEARCH DIRECTIONS

Due to the specific architecture of deep learning models, differential privacy faces several issues in the case of combining with these models. On the one hand, robust privacy protection framework is expected to protect all kinds of deep learning tasks; on the other hand, we hope that the model can still maintain high performance after perturbation. Considering the current challenges and existing solutions, the future research directions in this field may focus on the following three aspects: the evolution of differential privacy concepts in deep learning scenarios, the quantification of differential privacy protection, and the correlation between differential privacy and model robustness.

A. EVOLUTION OF DIFFERENTIAL PRIVACY RELATED CONCEPTS IN DEEP LEARNING

From privacy protection for traditional database to training datasets in deep learning, Differential privacy need to be developed to match multiple iterations of high-dimensional data in deep learning. Some practices can be taken include more precise privacy budget allocation or relaxation methods to enhance its practicality.

The relaxation of differential privacy has evolved with the expansion of its application. The most common relaxations, the (ε, δ)-DP, are those algorithms rely on Gaussian noise mechanisms or the privacy analysis following the composition theorems. Compared to the standard definition, (ε, δ)-DP provides asymptotically smaller privacy losses under composition and allows for greater flexibility in privacy protection mechanism choosing. Compared with (ε, δ)-DP, R'enyi differential privacy [53] and Concentrated Differential Privacy (CDP) [37] intuitively combine privacy budget with advanced composition theorems. Bun and Steinke [66] also relaxed the definition of differential privacy by approximating the linear upper bound of the moment function.

B. QUANTIFICATION OF PROTECTION PROVIDED BY DIFFERENTIAL PRIVACY

The protection of differential privacy is based on strict mathematical proof, but because of lack of intuitionistic interpretability, it is often questioned. In recent years, various attacks steal the privacy of training samples and against model robustness. It is challenging to determine the exact risk of an attacker re-identifying or reconstructing data under differential privacy. Some methods use the accuracy of membership inference attack and F1 score [67] as the evaluation indexes. Reiter *et al.* [68] assessed the privacy risk of data publishing by inference on synthetic data sets, but it is infeasible to run a set of inference attacks to estimate the risk before data publishing. If a model designer can provide the evidences of effective protection of differential privacy on both theoretical proof and experimental results, it will be an attractive security solution for deep learning applications.

C. RELATIONSHIP BETWEEN ROBUSTNESS OF DEEP LEARNING AND DIFFERENTIAL PRIVACY

Nowadays, differential privacy is mainly used to solve the privacy security problem of deep learning training data, and it only has a certain defense effect against membership inference attacks. In fact, although the threats that deep learning faces in practical applications are various and complex, they have the common feature, aiming at over-fitting of deep learning.

Over-fitting is an important but not the only reason of privacy leakage in machine learning models. It is an inherent problem in machine learning, which limits the prediction accuracy and generalization ability of the model. There is evidence that differential privacy in very large data sets can even prevent over-fitting to reduce prediction errors [69]. It means that machine learning and privacy researches may not always play a zero-sum game between utility and privacy, but have similar goals. Lécuyer *et al.* [70] pointed out that differential privacy can be used as a good means to against adversarial examples [71], which also expands the application area of differential privacy.

VI. CONCLUSION

At present, the research of differential privacy in the field of deep learning is still in its infancy. We want to figure out that the application of differential privacy mechanism in deep learning is the compromise of reducing availability or the win-win of availability and security. This paper gives some inspirations to some extent. We start with the threats which deep learning faced, demonstrate the related concepts of differential privacy in deep learning and summarize the characteristics of various models based on the deployment location of differential privacy. Finally, we summarize some key issues in this field, and suggest the direction for further research.

REFERENCES

[1] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.

[2] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2014, pp. 701–710.

[3] T. Zhou, S. Yang, L. Wang, J. Yao, and G. Gui, "Improved cross-label suppression dictionary learning for face recognition," *IEEE Access*, vol. 6, pp. 48716–48725, 2018.

[4] J. Pan, Y. Yin, J. Xiong, W. Luo, G. Gui, and H. Sari, "Deep learning-based unmanned surveillance systems for observing water levels," *IEEE Access*, vol. 6, pp. 73561–73571, 2018.

[5] D. Chicco, P. Sadowski, and P. Baldi, "Deep autoencoder neural networks for gene ontology annotation predictions," in *Proc. ACM BCB*, 2014, pp. 533–540.

[6] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.

[7] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive MIMO for hybrid precoding," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3027–3032, Mar. 2019. doi: 10.1109/TVT.2019.2893928.

[8] T. Ching *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *J. Roy. Soc. Interface*, vol. 15, no. 141, 2018, Art. no. 20170387.

[9] W. Huang and J. W. Stokes, "MtNet: A multi-task neural network for dynamic malware classification," in *Proc. Int. Conf. Detection Intrusions Malware, Vulnerability Assessment (DIMVA)*, vol. 9721, 2016, pp. 399–418.

[10] *The European General Data Protection Regulation (GDPR)*. Accessed: Mar. 1, 2019. [Online]. Available: https://gdpr-info.eu/

[11] T. Dalenius, "Towards a methodology for statistical disclosure control," *Stat. Tidskrift*, vol. 15, nos. 429–444, pp. 1–2, 1977.

[12] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[13] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," in *Proc. Int. Conf. Data Eng. (ICDE)*, Atlanta, GA, USA, Apr. 2006, pp. 1–24.

[14] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.

[15] R. Chi-Wing, J. Li, A. W.-C. Fu, and K. Wang, "(a, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 754–759.

[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography* (Lecture Notes in Computer Science), vol. 3876, S. Halevi and T. Rabin, Eds. Berlin, Germany: Springer, 2006, pp. 265–284.

[17] Apple. *S IOS10*. Accessed: Mar. 1, 2019. [Online]. Available: https://developer.apple.com/videos/wwdc2016/

[18] O. C. Novac, M. Novac, C. Gordan, T. Berczes, and G. Bujdosó, "Comparative study of Google Android, Apple iOS and microsoft windows phone mobile operating systems," in *Proc. 14th Int. Conf. Eng. Mod. Electr. Syst. (EMES)*, Jun. 2017, pp. 154–159.

[19] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2014, pp. 1054–1067.

[20] G. Fanti, V. Pihur, and Ú. Erlingsson, "Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries," *Proc. Privacy Enhancing Technol.*, vol. 2016, no. 3, pp. 41–61, 2016.

[21] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, and J. Shin. (2016). "Collecting and analyzing data from smart device users with local differential privacy." [Online]. Available: https://arxiv.org/abs/1606.05053

[22] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang. (2017). "Privacy loss in Apple's implementation of differential privacy on MacOS 10.12." [Online]. Available: https://arxiv.org/abs/1709.02753

[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[24] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," *Int. J. Secur. Netw.*, vol. 10, no. 3, pp. 137–150, 2015.

[25] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 3–18.

[26] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[27] B. Hitaj, G. Ateniese, and F. Pérez-Cruz, "Deep models under the GAN: Information leakage from collaborative deep learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2017, pp. 603–618.

[28] F. Doshi-Velez and B. Kim. (2017). "Towards a rigorous science of interpretable machine learning." [Online]. Available: https://arxiv.org/abs/1702.08608

[29] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *Proc. 23rd USENIX Secur. Symp.*, 2014, pp. 17–32.

[30] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2015, pp. 1322–1333.

[31] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *Proc. 25th USENIX Secur. Symp.*, 2016, pp. 601–618.

[32] C. Dwork, "A firm foundation for private data analysis," *Commun. ACM*, vol. 54, no. 1, pp. 86–95, Jan. 2010.

[33] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2016, pp. 308–318.

[34] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. 25th Int. Conf. Theory Appl. Cryptograph. Techn. (EUROCRYPT)*, 2006, pp. 486–503.

[35] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th IEEE Symp. Found. Comput. Sci. (FOCS)*, 2007, pp. 94–103.

[36] F. McSherry, "Privacy integrated queries," *Commun. ACM*, vol. 53, no. 9, p. 89, Sep. 2010.

[37] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.

[38] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 289–296.

[39] G. Acs, L. Melis, C. Castelluccia, and E. De Cristofaro, "Differentially private mixture of generative neural networks," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Jul. 2017, pp. 715–720.

[40] D. Su, J. Cao, N. Li, E. Bertino, M. Lyu, and H. Jin, "Differentially private k-means clustering and a hybrid approach to private optimization," *ACM Trans. Privacy Secur.*, vol. 20, no. 4, pp. 1–33, Oct. 2017.

[41] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted Boltzmann machines for collaborative filtering," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, vol. 19, nos. 1–2, pp. 791–798.

[42] D. P. Kingma and M. Welling. (2013). "Auto-encoding variational Bayes." [Online]. Available: https://arxiv.org/abs/1312.6114

[43] X. Zhang, S. Ji, and T. Wang. (2018). "Differentially private releasing via deep generative model (technical report)," [Online]. Available: https://arxiv.org/abs/1801.01594

[44] B. K. Beaulieu-Jones *et al.* (2018). *Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing*. [Online]. Available: https://www.biorxiv.org/content/10.1101/159756v5.abstract

[45] A. Triastcyn and B. Faltings. (2018). "Generating differentially private datasets using GANs." [Online]. Available: https://arxiv.org/abs/1803.03148v1

[46] V. Bindschaedler, R. Shokri, and C. A. Gunter, "Plausible deniability for privacy-preserving data synthesis," *Proc. VLDB Endowment*, vol. 10, no. 5, pp. 481–492, 2017.

[47] V. Bindschaedler and R. Shokri, "Synthesizing plausible privacy-preserving location traces," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 546–563.

[48] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, 2015, pp. 1310–1321.

[49] M. Abadi *et al.*, "On the protection of private information in machine learning systems: Two recent approches," in *Proc. IEEE 30th Comput. Secur. Found. Symp. (CSF)*, Aug. 2017, pp. 1–6.

[50] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *Proc. IEEE 51st Symp. Found. Comput. Sci.*, Oct. 2010, pp. 51–60.

[51] N. Phan, Y. Wang, X. Wu, and D. Dou, "Differential privacy preservation for deep auto-encoders: An application of human behavior prediction," in *Proc. 30th Conf. Artif. Intell. (AAAI)*, 2016, pp. 1309–1316.

[52] N. H. Phan, Y. Wang, X. Wu, and D. Dou, "Differential privacy preservation for deep auto-encoders: An application of human behavior prediction," *Mach. Learn.*, vol. 106, nos. 9–10, pp. 1681–1704, 2017.

[53] I. Mironov, "Rényi differential privacy," in *Proc. IEEE 30th Comput. Secur. Found. Symp. (CSF)*, Aug. 2017, pp. 263–275.

[54] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett, "Functional mechanism: Regression analysis under differential privacy," *Proc. VLDB Endowment*, vol. 5, no. 11, pp. 1364–1375, Jul. 2012.

[55] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[56] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[57] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.

[58] G. B. Arfken, H. J. Weber, and D. Spector, "Mathematical methods for physicists, 4th ed.," *Phys. Today*, vol. 20, no. 5, p. 79, 1996.

[59] E. Passow and T. J. Rivlin, "Chebyshev polynomials: From approximation theory to algebra and number theory," *Math. Comput.*, vol. 58, no. 198, p. 859, Apr. 1992.

[60] N. Phan, X. Wu, H. Hu, and D. Dou, "Adaptive laplace mechanism: Differential privacy preservation in deep learning," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 385–394.

[61] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *J. Mach. Learn. Res.*, vol. 12, pp. 1069–1109, Mar. 2011.

[62] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *Proc. IEEE 55th Symp. Found. Comput. Sci.*, Oct. 2014, pp. 464–473.

[63] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Localized rademacher complexities," in *Computational Learning Theory*. Cambridge, MA, USA: MIT Press, 2002, pp. 44–58.

[64] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," presented at the 5th Int. Conf. Learn. Represent., 2017.

[65] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson. (2018). "Scalable private learning with PATE." [Online]. Available: https://arxiv.org/abs/1802.08908

[66] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Proc. Theory Cryptogr. (TCC)*, vol. 9985, 2016, pp. 635–658.

[67] M. A. Rahman, T. Rahman, R. Laganière, N. Mohammed, and Y. Wang, "Membership inference attack against differentially private deep learning model," *Trans. Data Privacy*, vol. 11, no. 1, pp. 61–79, 2018.

[68] J. P. Reiter, Q. Wang, and B. Zhang, "Bayesian estimation of disclosure risks for multiply imputed, synthetic data," *J. Privacy Confidentiality*, vol. 6, no. 1, pp. 17–33, 2014.

[69] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "The reusable holdout: Preserving validity in adaptive data analysis," *Science*, vol. 349, no. 6248, pp. 636–638, 2015.

[70] M. Lécuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. (2018). "On the connection between differential privacy and adversarial robustness in machine learning." [Online]. Available: https://arxiv.org/abs/1802.03471v1

[71] I. J. Goodfellow, J. Shlens, and C. Szegedy. (2014). "Explaining and harnessing adversarial examples." [Online]. Available: https://arxiv.org/abs/1412.6572

JINGWEN ZHAO received the B.S. degree in information security from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China, in 2017, where she is currently pursuing the M.S. degree in information security. Her research interests include data security, differential privacy, and its applications in deep learning.

YUNFANG CHEN received the Dr. Eng. degree in computer science from Soochow University, Suzhou, China, in 2008. He is currently an Associate Professor with the School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests include big data processing and machine learning.

WEI ZHANG (M'19) received the Dr. Eng. degree in computer science from Soochow University, Suzhou, China, in 2008. From 2008 to 2011, he was a Postdoctoral Research Fellow with the Nanjing University of Posts and Telecommunications, Nanjing, China. He was a Visiting Scholar at Purdue University, in 2016. Since 2013, he has been a Professor with the School of Computer Science, Nanjing University of Posts and Telecommunications. His current research interests include computer version and machine learning.