Journal of Information Hiding and Privacy Protection
DOI:10.32604/jihpp.2020.010780
Review

A Survey on Recent Advances in Privacy Preserving Deep Learning

Siran Yin[1,2], Leiming Yan[1,2,*], Yuanmin Shi[1,2], Yaoyang Hou[1,2] and Yunhong Zhang[1,2]

[1]Nanjing University of Information Science & Technology, Nanjing, 210044, China
[2]Engineering Research Center of Digital Forensics, Ministry of Education, China
[*]Corresponding Author: Leiming Yan. Email: lmyan@nuist.edu.cn
Received: 13 August 2020; Accepted: 01 October 2020

Abstract: Deep learning based on neural networks has made new progress in a wide variety of domain, however, it is lack of protection for sensitive information. The large amount of data used for training is easy to cause leakage of private information, thus the attacker can easily restore input through the representation of latent natural language. The privacy preserving deep learning aims to solve the above problems. In this paper, first, we introduce how to reduce training samples in order to reduce the amount of sensitive information, and then describe how to unbiasedly represent the data with respect to specific attributes, clarify the research results of other directions of privacy protection and its corresponding algorithms, summarize the common thoughts and existing problems. Finally, the commonly used datasets in the privacy protection research are discussed in this paper.

Keywords: Deep learning; privacy preserving; adversarial learning; differentia-lly private

1 Introduction

In recent years, deep learning technology has been widely used in life, and related algorithms are emerging one after another, its superiority is constantly revealed. However, while enjoying the convenience provided by deep learning, people find that it is not without drawbacks, and the problem of privacy protection is particularly prominent. Deep learning is a technique for learning the intrinsic connection and representation level of training samples. The text samples and image samples are used to train deep learning model may have sensitive information. People using training samples can encode these samples into model parameters or even store the entire data set [1]. People do not know whether the samples provided for deep learning has been leaked, they only have little control over their personal data. For example, the neural network can calculate the credit risk from the personal attribute value [2,3], people also can use deep learning to enhance the semantics of Weibo to achieve the purpose of sentiment analysis [4].

The most common approach to protecting privacy is to use differential privacy technology (DP) to unbiasedly represent specific attributes of the data, DP refers to the samples with the most difference between the two input sample sets, which possesses theoretical rigor and without regard to the background knowledge of the attacker. Due to the complexity of the internal structure of deep learning, Song et al. expressed that the stochastic gradient descent algorithm (SGD) can be used to reduce the computational difficulty caused by DP by using the paradigm gradient cropping batching [5]. When the update packets in the SGD form a “minibatch” [6], the algorithm robustness can be greatly improved without increasing the computational cost. DP is suitable for privacy protection in deep learning [7], the privacy protection of text based on this method will be introduced in Section 3.3.

Privacy issues in machine learning has been everyone's attention over last decade, but until 2015, a more effective method to reduce the loss of privacy in deep learning was first proposed by Shokri et al. [8], the proposed method advocates selective sharing of sample parameters during deep learning, the shared sample parameters in method are controllable and the initiative is in the sample providers. This method is not only not affected by the algorithm of the specific task building model, but also can strike a balance between the accuracy of deep learning and the number of samples, which was a big step forward in the deep learning of privacy protection, after that, many people have optimized and diverged from this method.

Although image-based privacy protection has achieved a lot of results, there are still some shortcomings in the privacy protection of text analysis which contains a large amount of private information. The research in using deep learning method to protect text privacy will be a hot topic in the future. Therefore, this paper summarizes the text privacy protection, especially those achievements of text privacy protection based on deep learning. The basic framework of this paper is as follows; Section 2 briefly introduces the methods for text privacy attacks in deep learning. Section 3 focuses on the research progress. Section 4 discusses the challenges and prospects for privacy protection technologies. Section 5 is a summary of the full text.

2 Text Privacy Attack

2.1 Attacks on the Privacy of the Text Itself

For the text itself, the language is highly diverse, and the attributes of text may vary greatly because of the different ages and circumstances of different authors. These differences may have a large impact on natural language processing (NLP) models trained from text, which can cause the results produced by the model to deviate significantly from standard results, attackers can use these enough clues to identify the authors, their gender, age, and other important attributes. Sensitive information contained in the text is often leaked inadvertently, AOL (American OnLine) search data has experienced a breach, in August 2006, it released a detailed search log of some users [9]. Although the user information is no longer visible, personally identifiable information can still be obtained from the log information. Other sources of user text which is applied to the field of deep learning algorithms, such as Twitter, email, etc., also have problems of mining personal privacy based on the characteristics of the system model. Here is an example of adversarial attack against NLP, the attacker eavesdropped on the text classifier, captures hidden information from the classifier and attempts to recover information about the input text.

2.2 Attacks on the Privacy of Shared Text

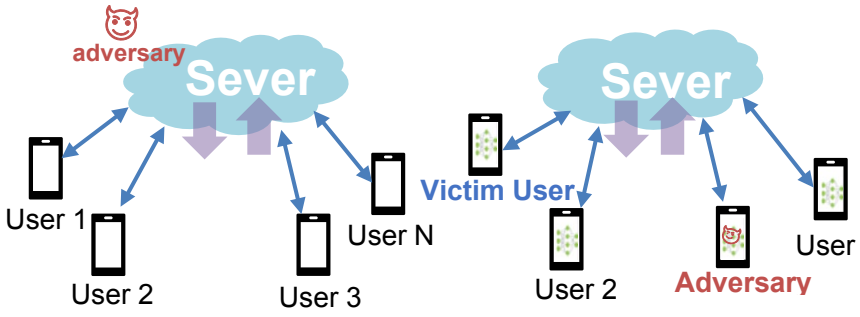

Figure 1: Two different ways of distributed deep learning training model. (a) Centralized learning. (b) Collaborative learning

For shared text data, privacy is often hidden in the text, such as: send samples to trusted third parties for NLP processing, or online translation of text, these samples possibly carry private information, enemies can use reverse engineering input to analyze sensitive information, especially when the enemy has a certain knowledge of the training model. Hitaj et al. used this idea, they proposed an attack method on collaborative deep learning [10], as shown in Fig. 1(a), the blue links show sharing of the model

parameters, if there are malicious participants in collaborative learning model, the centralized server is the only place that jeopardizes data privacy. Enemies train Generative Adversarial Networks (GAN) to generate prototype samples with the same distribution as the private target training set, while training samples, the malicious participants are always active and deceive the victim to release his private information. In Fig. 1(b), it is shown that any user can intentionally endanger any other user.

3 Privacy Protection Based on Text Deep Learning

3.1 K-Anonymity

While maintaining personal privacy, publishing data from tables containing personal records for analysis is an increasingly important issue today. In order to publish data about individuals without revealing sensitive information, Samarati proposed the k-Anonymity model [11] to protect privacy by requiring non-critical attributes to suppress or generalize leaked information, so that for each record in the modified table, at least k-1 other records have the same quasi-identifier value. However, the performance of k-Anonymity's most famous approximation algorithm depends linearly on the anonymity parameter k, its performance may be optimized.

Aggarwal et al. Proposed a new clustering method [12], which is a clustering method that anonymizes them before publishing them. Firstly, cluster the quasi-identifiers of the data records, and then publish the cluster center. To ensure the privacy of the data record, constraints are imposed, that is, each cluster must contain no less than a pre-specified number of data records. This technique is more versatile than k-Anonymity. Aggarwal et al. further studied r-Gather and the newly introduced clustering index called r-Cellular Clustering, and provided a constant factor approximation algorithm to explain this clustering, they also extended the algorithm to allow partial points to remain un-clustered. Therefore, by not releasing a small portion of the database record, it can be ensured that the data published for analysis has less distortion.

3.2 Homomorphic Encryption

Homomorphic encryption is a solution that solves the contradiction that users do not want to disclose information and the server needs to analyze data. Participants can encrypt the personal sensitive information and store it in the server, the server could process and analyzes the ciphertext, and participants will get messages that only participants can decrypt. The core of homomorphic encryption is the ability to perform operations directly on ciphertext, the result of the operation is decrypted and the decrypted result is the same as the clear text. This is the most direct and effective protection of participants' privacy [13].

Suppose there is an encryption function f such that $f(M)=M', f(N)=N'$, where M is clear text, encrypted to become ciphertext M', ciphertext N is the same, and decryption function $f-1$ exists which can decrypt the clear text before f encryption. Let $M'+N'=P'$, if $f-1$ decrypts P', the result is equal to the result of adding M and N, i.e., $f-1(p')=f-1(M'+N')=M+N$, then f is an encryption function that can perform homomorphic encryption.

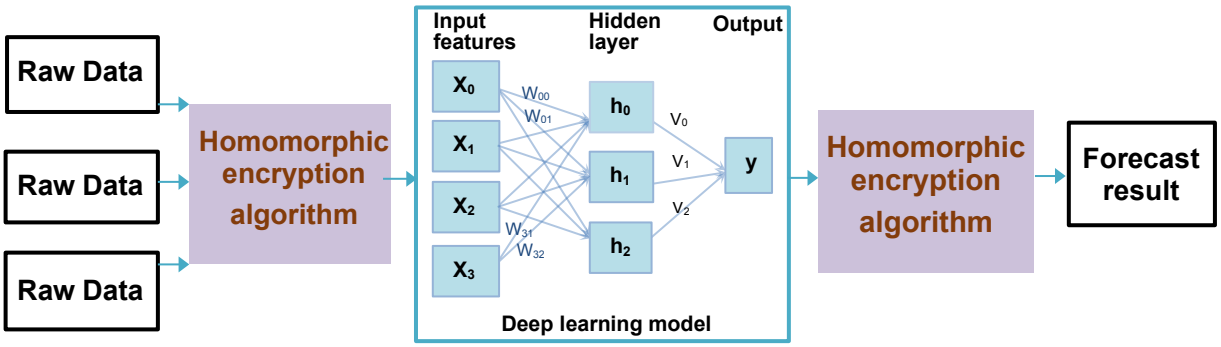

Figure 2: Homomorphic encryption based on deep learning application

3.3 Differential Privacy

Section 3.3 will focus on several deep learning methods for differential protection of privacy, which can also be used for text privacy protection. The differential privacy algorithm makes the training samples outputted by adjacent data sets not have significant differences, differential privacy can be deployed in three stages of deep learning, namely input layer, hidden layer and output layer, as shown in the Fig. 3. This section divides methods to protect privacy into these three stages, and then introduces them in categories.

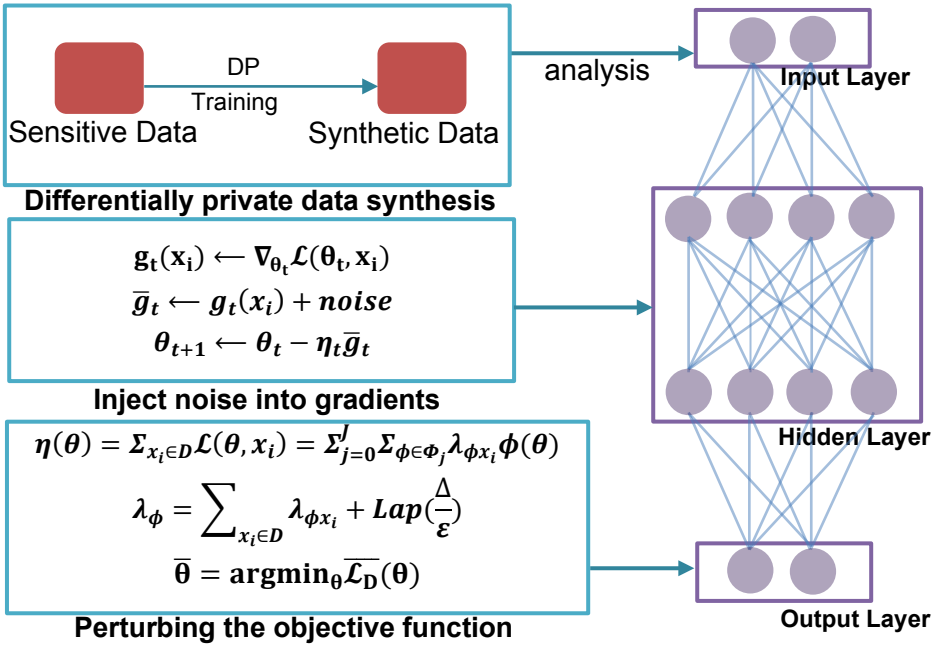

Figure 3: Applying differential privacy in three locations in deep learning model

3.3.1 Differential Privacy at the Input Layer

The differential privacy operation of the sample on the input layer is to preprocess the data so that the truly input samples contain the statistical characteristics of the original samples but the privacy information is protected, pre-processed samples can be used to generate models and used for analysis without revealing privacy.

Most pre-processing methods add noise to the sample model, for data clustering, Acs et al. proposed the k artificial neural network (ANN) [14], using Random Fourier Features (RFF) to transform data into low-dimensional space, then finds the required clusters using the core k-means algorithm [15] for data partitioning, and adds noise based on differential privacy based on random gradient descent (SGD) iteration. This method mainly focuses on optimizing the non-convex loss function, compared with the single generated model, the method proposed by Acs et al. has a shorter learning time, more detailed noise, and a higher sample availability. Based on differential privacy, Naoise Holohan et al. also proposed the use of generalized random response (RR) techniques [16], which is considered to be a difference of privacy under strict and loose conditions. RR enables managers to get distorted individual information, but can get an estimate of real information on this information, it selects the optimal RR mechanism by the value of

$$g(\varepsilon,\delta) = \frac{\delta(e^{\varepsilon} + \delta)}{(e^{\varepsilon} + 2\delta - 1)^2} \tag{1}$$

where $g: R \times R \to R$, $\delta > 0$ and $0 < \pi \le \frac{1}{2}$, the result depends on the known true value π.

When sensitive attributes are correlated with certain quasi-identifiers, Gaoming Yang et al. proposed stoically using the invariant random response method to interfere with strongly associated attributes [17], so that make the sample meet the requirement of local $\varepsilon-$difference privacy.

Define 1. $\varepsilon-$local differential privacy set multiple participants, each participant has a piece of data, and gives the privacy protection algorithm F and its definition domain $Dom(F)$ and value domain $Ran(F)$. If algorithm F obtains the same output result $l'(l' \in Ran(F))$ on any two customer records l and $l'(l,l' \in Dom(F))$, then F satisfies $\varepsilon-$local differential privacy:

$$P_r[F(l)=l^*] \le e^{\varepsilon} \times P_r[F(l')=l^*] \tag{2}$$

3.3.2 Differential Privacy in Hidden Layer

Adding noise to the hidden layer is a more common solution. This solution can protect the privacy content more intuitively. Since the hidden layer of deep learning is mostly iterative, the added noise operation will be easier to implement. Applying differential privacy to the hidden layer was also the earliest proposed solution. The original intention is to prevent the enemy from mastering the accurate personal information of the training through the output data. This solution has many innovative improvements, such as the more accurate added noise and the more stringent measurement privacy loss, these will be introduced later, all of which have significant implications for the differential privacy of deep learning.

Differential privacy stochastic gradient descent (dpSGD) is a basic method whose main operations are program disinfection and privacy accounting. Disinfection of the program means that the gradient of the sample is cropped to limit its sensitivity, and noise is added to each gradient in turn. The purpose of privacy accounting operation is to track the loss of privacy.

The addition of noise in the neural network distributed system under differential privacy will have a great impact on the samples. At the same time, the number of iterations of deep learning hidden layer will increase, resulting in high cost. To avoid the above problems, the number of noises can be more accurately controlled, the composition theorem can provide privacy guarantee for interactive query, but now most of the composition theorems of neural networks only provide a loose boundary for privacy output, which leads to the fast consumption of privacy budget in several iterations, so does the strong composition theorem [18], which has poor practicability. Abadi et al. proposed moments accountant [19], that is, the combination of the composition theorem and differential privacy to reduce the loss of privacy. Specifically, privacy accounting will interfere with sample data in batches, and each stage needs to satisfy $(O(qT\varepsilon),qT\delta)-DP$, the sampling rate q needs $q=L/n$. It can be seen from Tab. 1. that moments accountant is a better scheme. In particular, if $\delta = min\ \lambda exp\left(\alpha M(\lambda)-\lambda\varepsilon\right)$ is satisfied, we set the generating function of moment is αM, then the mechanism M satisfies $(\varepsilon,\delta)-DP$ is correct.

Table 1: Privacy budget restriction under different combination methods

Method	Naive composition	Strong composition	Moments accountant
Privacy budget bound	$(O(qT\varepsilon),qT\delta)-DP$	$(O(q\varepsilon\sqrt{T\log\frac{1}{\delta}}),qT\delta)-D$	$(O(q\varepsilon\sqrt{T}),\delta)-DP$

Based on the selective random gradient descent, Shokri et al. proposed a neural network distributed system with differential privacy [8], which provided new ideas for text privacy protection on the mobile side. This scheme uses the sparse vector technique to apply differential privacy to parameter updates, which can alleviate the loss of privacy associated with parameter selection and shared parameter values. Participants can independently train on their data sets and selectively share some parameters during training. The control of these parameters is in the hands of the participants themselves, which allows

participants to retain their privacy and benefit from models by other participants.

Applying Concentrated Differential Privacy (CDP) to privacy accounting to achieve a strict estimate of privacy loss is another attempt by Yu et al. [20], who implemented a dynamic privacy budget allocator during the training process to improve the accuracy of the model. It turns out that their program has effectively improved privacy loss accounting, training model efficiency and quality under a given privacy budget.

The differential private convex optimization algorithm [21] is suitable for real-world situations and provides privacy and practical guarantees. The general convex optimization "disturbs" the objective function by adding random linear terms and releasing the minimum value of the perturbed target, but only in when the output of the mechanism happens to be the minimum value of the interfered target, the target interference can provide privacy guarantee. It is stated that privacy and utility guarantees can be obtained even if the system releases the noisy "approximate" minimum of the perturbed target.

3.3.3 Differential Privacy at the Output Layer

In the output layer, the loss function is usually used to contrast the dissimilarity between the foretell value of $\overline{y}$ (true value) and $\overline{y} = f(X)$ in deep learning. The smaller the loss function, the stronger predictive power of the model on the training samples, the greater the output impact on the loss function, the algorithm is more optimized.

Different from the previous two sections, functional mechanism (FM) [22] can protect privacy by adding noise to the objective function. There are many restrictions on the direct disturbance of the output results, which can only be performed on the standard types of regression analysis. In general, the FM injects Laplace noise into the approximation coefficient of the objective function polynomial $\overline{F_D(\omega)}$, and then uses the model parameter $\overline{\omega}$ to minimize the objective function $\overline{F_D(\omega)}$.

Phan et al. proposed a scheme to increase adaptive noise to reduce the cost of adding noise [23], which interferes with the affine transformation of neurons and loss functions, and injects noise adaptively into the features according to the impact weight of each output on the result, where the lighter the feature weight adds more noise. The model has the generality of deep learning, because it adds perturbations to features, affine transformation layers, and loss functions based on differential privacy, and the practicality of the model is reflected in ensuring that the privacy budget is not related to the number of training periods.

3.4 Adversarial Learning

Adversarial training refers to the method of constructing adversarial samples in the training process of the model and mixing the adversarial samples with the original samples to train the model. It can be said that different methods of adversarial attack determine different methods of adversarial training.

In order to satisfy the two features of adversarial disturbance, that is, the disturbance is difficult to observe and the added disturbance must have the ability to make model output different from original output, previous studies have shown how to generate effective adversarial disturbance. Unlike the image domain, the perturbation added in the text can be discrete or continuous, generally speaking, discrete perturbation refers to minor modifications to input text characters directly, and continuous perturbation refers to the disturbance directly added to the word vector matrix in the input text [24]. The discrete disturbances in the text processing are shown in the Fig. 4.

Part (1) refers to the original input text, and part (2) refers to the text after discrete perturbation, although only a few characters have been modified but the model produces a completely different output. In text processing, the feature of adversarial disturbance requires that the semantic consistency between the adversarial samples and the original sample, that is, the added disturbance should not change the semantic of the original sentence as much as possible.

Traditional methods to protect privacy (such as adding interference to samples) will have a negative effect on the accuracy of analysis privacy in varying degrees, and differential privacy methods cannot

protect text privacy in adversarial attacks, the above situation can be avoided by using adversarial learning [25] proposed by Ganin et al.

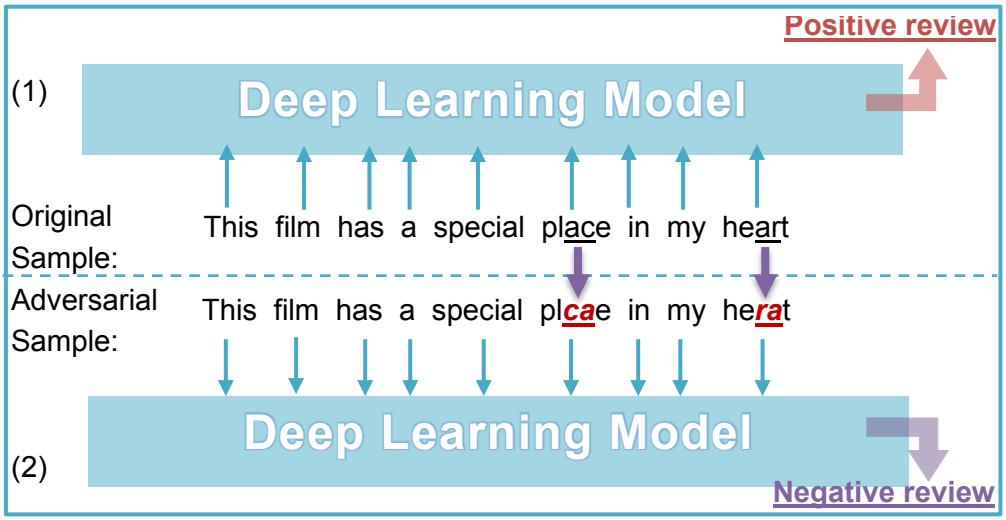

Figure 4: The adversarial disturbance of text processing in deep learning

Assuming that h cannot only accurately predict the result y but also represent the attribute, let x represents the cross-entropy function, then we can get

$$\hat{\theta} = \min_{\theta_M} \max_{\theta_D} X(\hat{y}(x;\theta_M), y) - \lambda \cdot X(\hat{b}(x;\theta_D), b) \tag{3}$$

The negative sign of the second term means counter loss, which can be realized by gradient inversion layer during back propagation, it is worth noting that the linear output component is trained as a good predictor, but h is trained as the best for the main task and the worst for the auxiliary task.

Coavoux et al. [26] proposed an adversarial scenario: Attackers eavesdrop on the hidden representation of input language samples of users, and attempt to recover the original text content. Coavoux et al. choose the standard LSTM architecture to calculate the fixed size representation $r(x)$ from a series of tags $x = (X_1, X_2 \ldots, X_n)$ and project it into the embedding space, the construction parameters /r of r include LSTM parameters and word embedding.

$$L_m(\theta_\gamma, \theta_p) = \Sigma_{i=1}^{N} - \log P(y^{(i)} \mid x^{(i)}; \theta_\gamma, \theta_p) \tag{4}$$

$r(x)$ is the output of the encoder, and it has the parameter u_0 used to predict the label y of the text. Coavoux et al. trained the model to reduce the possibility of label y becoming negative logarithm, this classifier optimizes the balance between the utility and privacy of neural representation, and has low computational cost, this scheme also has generality, because the generation of confrontation cost does not need to specify private variables. However, as a kind of adversary learning scheme, its shortcomings are also very obvious, under the adversary attack which does not meet the expectation, the specific adversary learning defense scheme shows the vulnerability.

NLP mainly deals with logical language, which is a sub meadow of artificial intelligence. The language NPL uses often provided by users themselves, which inevitably contains privacy information, the existence of privacy information is not directly visible, but hidden between the lines, moreover, due to people of different ages and regions have different writing habits, the differences in language characteristics will be great. We assume that sensitive attributes do not supervise the text representation, but exist in the vector representation of the text, Li et al. proposed the privacy of unsupervised image

representation [27]. This model can measure the privacy by comparing the peak value of the user's image and the reconstructed image of the attacker, and find a balance between the accuracy and privacy of the image. The storage problem of personal information in data set solved by Carlini et al. [28] is quite similar to the privacy represented by unsupervised image. Assuming that the attacker has obtained a trained language samples, they will combine and analyze some statements, but most of analysis results are wrong. Let's consider in another direction: when attackers can access the hidden layer of deep learning, they will try to obtain the language samples of input. In view of this situation, Li et al. proposed a training scheme based on Gan method to improve the robustness of deep learning [29], which may overestimate privacy, because it assumes that the number of attackers is only one, and when the parameters of the main model are fixed, the classifier will be retrained to evaluate privacy.

4 The Problems and Development Trend of Deep Text Learning in Privacy Protection

Due to the specific architecture of deep learning models, privacy protection algorithms show many problems when used in combination with them. On the one hand, we hope that the privacy of deep learning samples can be truly and effectively protected. On the other hand, the deep learning's high performance is something we do not want to weaken after adding privacy protection.

For privacy protection using homomorphic encryption method, it has great limitations, for example, it only supports integer samples and needs fixed multiplication depth, cannot add and multiply indefinitely. Fully-homomorphic encryption does not support comparison and operation of taking the maximum. The most important thing is that homomorphic encryption technology has too much operation cost, which consumes a lot of computing resources in combination with its own deep learning algorithm, the algorithm performance will be greatly reduced.

For the application of differential privacy in text deep learning, it is based on strict mathematical proof, but lack of evidence on both theoretical and experimental results, proving that differential privacy in deep learning can effectively protect privacy. In recent years, there are some attacks that steal the privacy of data sets and destroy the robustness of the model, Reiter et al. [30] evaluated the privacy risks of data publishing by training synthetic datasets, in order to run a set of inference attacks and make inferences before data is published, but it is not feasible to estimate the risk before data release, it is challenging to determine the risk of privacy re-identification by attackers in different data sets. Over fitting is an inherent problem of machine learning, which may causes privacy leakage and limits the model's generalization capacity and prediction accuracy. There is evidence that if the data set is large enough, differential privacy can prevent over fitting to reduce prediction error [31], which means that deep learning and privacy protection could not always seek a balance between training performance and privacy issues.

For adversary learning, its attack is threatening, but there are few defense methods [32], most of the existing work has its own limitations, such as application scenarios, constraints and the problems of method itself. The poor portability of the classifier to correctly identify the adversary samples and adversary instances is also the reason why adversary learning is not easy to apply in deep learning. In addition, the adversary learning of text lacks a benchmark and open-source Toolbox: people put forward various methods to study adversary attack and defense in deep learning of text, but there is no benchmark to show that which measurement standard is better in some cases, one of the important reasons is that a measurement method performs well in this work, but it is ineffective in other work. If we use the open-source toolbox, people can easily do further research, so as to reduce the repeated time consumption and promote the research and development in this field. In the text field, the visual analysis framework proposed by Laughlin and others [33], if it can be combined with the open-source toolbox (such as the image field Papernot [34]), we will obtain more diversified attack and defense means.

Table 2: Comparison of different privacy protection methods in deep learning

Research	Main advantage	Main shortcoming	Basic method
Aggarwal et al. [12]	1.Allow samples to not cluster 2. Small sample distortion	Vulnerable to background knowledge attack	k-Anonymity
Phonget al. [13]	Combining homomorphic encryption with deep learning	Cannot applicable to shared data	Homomorphic encryption
Abadi et al. [19]	Limits on loss of privacy are stricter	Not suitable for shallow models and small data sets	Differential privacy
Shokri et al. [8]	1. Suitable for shared data 2.Prevent overfitting in some cases	1. Slow learning speed 2. Not available for local SGD	
Acs et al. [14]	1. Suitable for high-dimensional large data sets 2. High data availability	Weak privacy protection for multiple hidden layer models	
Holohan et al. [16]	Differential privacy with strict and relaxed conditions is considered.	Not available for multivalued responses	
Yang et al. [17]	1.Different post random disturbance methods are designed to protect privacy 2.High utilization of data		
Yu et al. [20]	Less privacy loss in a single iteration	Model accuracy is sensitive to noise scale	
Iyengar et al. [21]	It can be used without any super parameter adjustment	High- dimensional kernel learning is not considered	
Zhang et al. [22]	The expansion of Laplasse mechanism requires no assumption of attack	Only for standard types of regression analysis	
Phan et al. [23]	Privacy budget regardless of the number of training cycles in some cases	Unable to easily migrate to other deep learning models	
Wang et al. [24]		It should not be used in datasets other than English	Adversarial learning
Yaroslav et al. [25]	1.Suitable for document emotion analysis and image classification 2. High accuracy of analysis privacy	Defensive fragility in adversarial attacks that do not meet expectations	
Coavoux et al. [26]	1.Low calculation cost 2.High universality	Failure to defend against undesired attacks	
Li et al. [29]	Supports high-quality NLP inference while transmitting privacy samples	Model confidentiality has not been proved by theory	

5 Conclusion

At present, the research of privacy protection in the field of text deep learning is still in its infancy. We want to know a win-win compromise between availability and security of deep learning, this paper provides some ideas in these respects. First, we start with the threat of privacy in deep learning, and classified introduce several attack methods according to sample status. Then we introduce the methods of combining k-anonymity, homomorphic encryption, differential privacy and adversarial learning in deep learning to achieve the purpose of protecting privacy. Compared with the traditional privacy protection methods, the privacy protection algorithm based on deep learning can make use of deep learning technology by integrating various types of multisource heterogeneous data, adding more precise noise or fuzzy attribute differences for privacy. Currently, there are few experimental methods for text privacy only, but the methods described above can be used as the basis for the research of text privacy information protection in deep learning. Finally, we summarize some existing problems and further research directions in this field.

Funding Statement: This work is supported by the NSFC [Grant Nos. 61772281, 61703212, 61602254]; Jiangsu Province Natural Science Foundation [Grant No. BK2160968]; the Priority Academic Program

Development of Jiangsu Higher Education Institutions (PAPD) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET).

Conflicts of Interest: We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals, "Understanding deep learning requires rethinking generalization," *CoRR*, abs/1611.03530, 2016.

[2] D. D. Wu, D. L. Olson and C. Luo, "A decision support approach for accounts receivable risk management," in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 12, pp. 1624-1632, 2014.

[3] I. S. Duma, B. Twala and T. Marwala, "Predictive modeling for default risk using a multilayered feedforward neural network with Bayesian regularization," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, pp. 1–10, 2013.

[4] Y. X. He, S.T.Sun, F. F. Niu and F. Li, "An emotional semantic enhanced deep learning model for weibo emotional analysis," *Chinese Journal of Computers*, no. (4), 2013 (in Chinese).

[5] S. Song, K. Chaudhuri and A. Sarwate, "Stochastic gradient descent with differentially private updates," in *2013 IEEE Global Conf. on Signal and Information Proc.*, pp. 245–248.

[6] O. Dekel, R. GiladBachrach, O. Shamir and L. Xiao, "Optimal distributed online prediction using minibatches," *Journal of Machine Learning Research*, vol. 13, pp. 165–202, 2012.

[7] K. Chaudhuri and C. Monteleoni, "Privacypreserving logistic regression," in *Proc. of NIPS*, USA, 2008.

[8] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," *2015 53rd Annual Allerton Conf. on Communication, Control, and Computing (Allerton*), Monticello, IL, 2015, pp. 909–910.

[9] G. Pass, A. Chowdhury and C. Torgeson, "A picture of search," in *The First Int. Conf. on Scalable Information Systems*, vol. 152, 2006.

[10] B. Hitaj, G. Ateniese and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security*, ACM. 2017.

[11] P. Samarat, "k-Anonymity," in van Tilborg H.C.A., Jajodia S. (eds), *Encyclopedia of Cryptography and Security*. Springer, Boston, MA, 2011.

[12] G. Aggarwal, R. Panigrahy, T. Feder, D. Thomas, K. Kenthapadi *et al.,* "Achieving Anonymity via Clustering," *ACM Transactions on Algorithms*, 2010.

[13] L. Phong, Y. Aono, T. Hayashi, L. Wang and S. Moriai, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Transactions on Information Forensics and Security*, pp. 1.

[14] G. Acs, L. Melis, C. Castelluccia and E. De Cristofaro, "Differentially Private Mixture of Generative Neural Networks," in *2017 IEEE Int. Conf. on Data Mining (ICDM)*, New Orleans, LA, 2017, pp. 715–720.

[15] B. Schölkopf, A. J. Smola and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, 1998.

[16] N. Holohan, J. L. Douglas and O. Mason, "Optimal differentially private mechanisms for randomised response," *IEEE Transactions on Information Forensics and Security*, arXiv:1612.05568, pp. 2726–2735, 2017.

[17] G. M. Yang, H. M. Zhu, X. J. Fang and S. Z. Su, "Invariant Post-Random Response Perturbation for Correlated Attributes Under Local Differential Privacy Constraint," *Acta Electronica Sinica*, vol. 47, no. 5, pp. 1079–1085, 2019 (in Chinese).

[18] C. Dwork, G. N. Rothblum and S. Vadhan, "Boosting and Differential Privacy," in *2010 IEEE 51st Annual Sym. on Foundations of Computer Science,* Las Vegas, NV, 2010, pp. 51–60.

[19] M. Abadi, A. Chu, I. Goodfellow and H. Brendan McMahan, "Deep learning with differential privacy," in *Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security*, 2016.

[20] L. Yu, L. Liu , C. Pu, M. E. Gursoy and S. Truex, "Differentially private model publishing for deep learning," in *IEEE Symposium on Security and Privacy*, 2019.

[21] R. Iyengar, J. P. Near, D. Song, O. Thakkar, A. Thakurta and L. Wang, "Towards Practical Differentially Private Convex Optimization," in *2019 IEEE Sym. on Security and Privacy (SP), San Francisco,* CA, USA, 2019, pp. 299-316.

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

[23] 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)*, pp. 385–394.

[24] W. Wang, L. Wang, B. Tang, R. Wang and A. Ye, "A survey on adversarial attacks and defenses in text," arXiv preprint arXiv:1902.07285, pp. 1–13. 2019.

[25] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle *et al.,* "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–35, 2016.

[26] M. Coavoux, S. Narayan and S. B. Cohen, "Privacy-preserving neural representations of text", arXiv preprint arXiv:1808.09408 2018.

[27] M. Li, L. Z. Lai, N. Suda, V. Chandra, D. Pan, "PrivyNet: A flexible framework for privacy-preserving deep neural network training with a fine-grained privacy control," arXiv:1709.06161, 2017.

[28] N. Carlini, C. Liu, J. Kos, Úlfar Erlingsson and D. Song, "The secret sharer: Measuring unintended neural network memorization & extracting secrets," *CoRR*, 2018.

[29] Y. Li, T. Baldwin and T. Cohn, "Towards robust and privacy-preserving text representations," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 25–30, Association for Computational Linguistics, Melbourne, Australia. 2018.

[30] 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.

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

[32] S. Zhang, X. Zuo and J. Liu, "Counter-sample problems in deep learning," *Chinese Journal of Computers*, no. 8, 2019 (in Chinese).

[33] B. Laughlin, C. Collins, K. Sankaranarayanan and K. El-Khatib, "A visual analytics framework for adversarial text generation," arXiv preprint arXiv:1909.11202, 2019.

[34] N. Papernot, I. Goodfellow, R. Sheatsley, R. Feinman and P. McDaniel, "Cleverhans v1.0.0: An adversarial machine learning library," arXiv preprint arXiv:1610.00768, 2016.