第50卷第8期 中国科学技术大学学报 Vol. 50, No. 8
2020年8月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Aug. 2020
文章编号:0253-2778(2020)08-1064-08

Differential privacy protection method for deep learning based on WGAN feedback

TAO Tao[1,2], BAI Jianshu[1], LIU Heng[1], HOU Shudong[1], ZHENG Xiao[1,2]
(1. *School of Computer Science and Technology, Anhui University of Technology, Ma'anshan* 243002, *China*;
2. *Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei* 20026, *China*)

Abstract: Aiming at the problem that attackers may steal sensitive information of the deep learning training dataset by some technological means such as the Generative Adversarial Network(GAN), combining the differential privacy theory, the differential privacy protection method was proposed for deep learning based on the Wasserstein generative adversarial network (WGAN) feedback parameter tuning. This privacy protection method is realized by optimization of the stochastic gradient descent, gradient clipping of setting gradient threshold, and noise adding to the optimization process of deep learning; WGAN was used to generate optimized results similar to the original data. The difference of the generated results and the original data were used for feedback parameter tuning. The experiment result shows that this method can effectively protect sensitive private information of the dataset and has preferable data usability.

Key words: differential privacy protection; deep learning; Wasserstein generative adversarial network(WGAN)
CLC number: TP18 **Document code**: A doi:10.3969/j.issn.0253-2778.2020.08.004
2010 Mathematics Subject Classification: 94B15

Citation: TAO Tao, BAI Jianshu, LIU Heng, et al. Differential privacy protection method for deep learning based on WGAN feedback[J]. Journal of University of Science and Technology of China, 2020,50(8):1064-1071.
陶陶,柏建树,刘恒,等. 基于WGAN反馈的深度学习差分隐私保护方法[J]. 中国科学技术大学学报,2020,50(8):1064-1071.

基于WGAN反馈的深度学习差分隐私保护方法

陶 陶[1,2],柏建树[1],刘 恒[1],侯书东[1],郑 啸[1,2]
(1.安徽工业大学计算机科学与技术学院,安徽马鞍山 243002;2.合肥综合性国家科学中心人工智能研究院,安徽合肥 230026)

摘要:针对攻击者可能通过某些技术手段如生成式对抗网络(GAN)等窃取深度学习训练数据集中敏感信息的问题,结合差分隐私理论,提出经沃瑟斯坦生成式对抗网络(WGAN)反馈调参的深度学习差分隐私保护的方法.该方法使用随机梯度下降进行优化,设置梯度阈值进行梯度裁剪,对深度学习的优化过程添加噪声实施隐私保护;利用WGAN生成与原始数据相似的最优结果,对比生成结果与原始数据的差异进行反馈调参.实验结果表明,该方法可以有效保护数据集的敏感信息并且具有较好的数据可用性.

关键词:差分隐私;深度学习;沃瑟斯坦生成式对抗网络(WGAN)

0 Introduction

As a hot branch of machine learning, deep learning has made rapid development since it was raised by Hintonet al[1]. Currently this kind of machine learning has shown a preferable performance far surpassing traditional methods in almost all of artificial intelligence sections. A large number of key breakthroughs have been made in many areas like the image, voice, and natural language processing. One of the important reasons for above achievements is the usability of large and representative dataset used in the neural training network. But, with the fast development of the deep learning, security and privacy problems have gained more and more concern.

Received: 2020-06-05; **Revised**: 2020-08-18
Foundation item: Supported by the Key Research and Development Program Project of Anhui Province of China(201904d07020020), the Natural Science Foundation Project of Anhui Province of China(1908085MF212, 2008085MF190, 1808085QF210), the Program for Synergy Innovation in the Anhui Higher Education Institutions of China(GXXT-2020-012).
Biography: TAO Tao(corresponding author), male, born in 1977, PhD/Associate Professor. Research field: Privacy protection and Deep learning. E-mail: taotao@ahut.edu.cn

In deep learning, privacy means that a person has the right to decide his personal data shall not be made public. The performance of deep learning has close relationship with the training dataset scale and diversity, while the data usually contains a lot of personal sensitive information, which can be obtained by attackers with methods of renewing part of deep learning model data or stealing the training model directly. How to increase the usability of training data without disclosing users' sensitive data has become the main problem to be resolved by deep learning.

In view of the related problems, scholars at home and abroad have done a lot of fruitful research. The earliest privacy protection model is k-anonymity raised by Sweeney et al[2]. After attackers obtain data processed by k-anonymity, they will get at least K differential data records thus unable to make accurate judgement. The raise of l-diversity[3] and t-closeness[4] all made improvement for k-anonymity. But this kind of data anonymity method can't guarantee strict privacy and the data usability reduction is usually caused by missing attributes. Differential privacy raised by Dwork et al[5] remedies the limitation of traditional privacy protection method. This model shows less concern for the background knowledge of attackers and makes strict definition and evaluation for the privacy protection degree. The combination of differential privacy and deep learning has become one the hot points of privacy protection study. In 2016, Abadi et al[6] put differential privacy technology into the process of deep learning training model. Privacy protection was realized by adding noise to optimization process, and real time accounting method was raised to evaluate the privacy information loss. Nicolas et al[7] raised the privacy aggregation method aiming at teacher model. This PATE model can provide more reliable protection for training data. Ian et al[8] raised a new deep learning model- Generative adversarial Network in 2014, through which preferable output results will be got by the antagonistic learning between the generated model and the discriminated model. In this way, similar data as the training dataset can be generated. Combining the unsupervised learning generative adversarial network and the convolutional neural network, AlecRadford et al[9] raised DCGAN model so as to get better expressive image features. But there is still the problem of the single oscillation and the privacy protection issue of unmarked data is also not concerned. Combining differential privacy and generative adversarial network for the first time, Beaulieu Jones et al[10] raised to generate medical data by using differential privacy to support classified generative adversarial network. Liyang Xie et al[11] proposed a privacy preserving generative adversarial network (DPGAN) that preserves the privacy of the training data in a differentially private sense . The difference from our work is that this method does not perform feedback adjustment of privacy parameters, and the correction of privacy parameters is mainly concentrated in the privacy budget. Most of the above methods combine differential privacy and traditional deep learning models to achieve privacy protection, mainly by adding random noise outside the training process. There is no strict quantitative analysis towards the noise amount added to the parameter. Too much noise will affect the usability of the output result, while too little noise will also affect the protection effect.

Based on the above information, this paper referred to the idea of Abadi and others, introduced the differential privacy theory into the deep learning model training and designed the privacy protection model for deep learning based on Wasserstein generative adversarial network feedback. In each training step, the training sample was extracted from dataset randomly, the gradient clipping threshold of the training model was set, gradient was adjusted and noise was added with iterative operation. We adjust the privacy parameters in the model through privacy feedback. This paper used Mnist dataset to verify the experiment. The result shows that compared with the traditional deep learning network, the accuracy rate with the WGAN training is relatively higher, and this method can guarantee privacy protection within a reasonable privacy budget.

The main contributions of this article are as follows:

(Ⅰ) The depth learning differential privacy algorithm is designed, the gradient clipping threshold is set, and the noise is added in the optimization process of deep learning parameters by combining the differential privacy theory, parameters affecting the noise are set in groups, and the gradient output is disturbed to realize privacy protection.

(Ⅱ) The defects of the traditional GAN in training are found, the Wasserstein distance in WGAN is proposed to solve the problems of instability and mode collapse in the traditional GAN. By using the best result similar to the original training data, the differences between the

WGAN result and the original data are compared. The adjusted parameters can satisfy the data privacy protection under the premise of ensuring the data availability.

1 Preliminaries

1.1 The theoretical basis of differential privacy

Differential privacy preserving is one kind of privacy protection technology based on data distortion for attackers with a strong knowledge background. It mainly adds noise to the real data or statistical results, interferes with the relevant sensitive data, and ensures that the data can still be used later. The differential privacy protection model can ensure that the output results will not be affected when a record is inserted or deleted centrally.

1.1.1 Definition of differential privacy

Definition 1.1 Given two neared-neighbor datasets D and D' that differ by at most one record, for a random algorithm M, its value range is Range(M). If any output of this algorithm in the above dataset meets the following conditions

$$\Pr[M(D) \in S] \leqslant e^{\varepsilon} \times \Pr[M(D') \in S] + \delta \tag{1}$$

The above random method provides the differential privacy of (ε, δ), ε is the privacy budget, indicating the degree of privacy protection. The smaller the ε value, the greater the interference to the output, so the more conducive to privacy protection. δ is the error value, indicating the probability of privacy disclosure.

1.1.2 Gaussian mechanism

The real-valued function which is similar to that with the differential privacy mechanism: $f: D \to R^d$.

Add a random noise to the output of f to protect the relationship between the data in the dataset and realize differential privacy protection. The volume of the added noise is directly proportional to the sensitivity of the output of f.

The global sensitivity of f is defined as follow

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

Among them, D and D' are the nearest neighbor datasets with a maximum difference of one record. Gaussian mechanism is used to add the Gaussian noise to the actual output value of f

$$f(D) = f(D') + N(0, S_f^2 \sigma^2) \tag{3}$$

where $N(0, S_f^2 \cdot \sigma^2)$ is a normal Gaussian with a mean of 0 and a standard deviation of $S_f^2 \cdot \sigma^2$.

Only when the algorithm is random, that is, when the output and distribution are related, will the data be replaced with appropriate noise protection, which will not have much effect on the result. If some data is interfered with or even deleted, its distribution does not change much, thus achieving the goal of differential privacy.

1.2 The foundation of the deep learning theory

Deep learning is a subfield of machine learning that can be supervised or unsupervised.

Deep learning is powerful because it uses basic concepts of data as nested concepts. In the nested concept hierarchy, simple concepts are refined to obtain complex concepts. Deep learning has been applied in many research fields, including computer vision, speech recognition and bioinformatics.

1.3 Generative adversarial networks

Due to the complexity of the deep learning network model, GAN is easy to remember the samples for the model training. When GAN is applied to sensitive data, the centralized distribution may lead to the leakage of some sensitive information. For A high-quality generated distribution, the privacy of the raw data can be protected by distributing the data only to the public or to restricted individuals, rather than to the raw data. But it is still possible to recover sample data by repeatedly sampling in the distribution. The attack model designed by Hitaj et al[12] can reconstruct the original training samples from the training samples.

The generative adversarial network[8] model is used to estimate the underlying distribution of the dataset and randomly generate actual samples according to its estimated distribution. Its basic idea is to use two mutually "game" models: a generation model G and a discrimination model D. The training goal of the generation model G is to generate results similar to the real data as much as possible. Through this generation result, the probability of judging errors by the discriminator is maximized, so that the discriminator mistakenly believes that the generated results are the real results of the data. The training goal of the discriminant model D is to maximize the discriminant accuracy of generated results and real results as much as possible. In the training process, the mutual "game" between G and D makes the performance of the two models enhanced at the same time. The objective function of GAN is

$$\min_G \max_D V(D,G) = E_{x \sim P_{\text{data}}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{4}$$

2 Method

2.1 Deep learning of differential privacy algorithms

The privacy protection method of the deep learning model is usually to protect the privacy of

the training data by processing the final parameters in the training process. We can treat the whole process as a black box. But, these parameters do not describe the dependence of training data effectively and strictly. Adding the over bottom noise to the parameters will make the selection according to the worst case, which will destroy the usability of the learning model. Therefore, it is necessary to design an effective method of differential privacy in the training process of deep learning.

Algorithm2.1 In the process of deep learning, the minimum loss function $L(\theta)$ is used to train the relevant parameter model, and the basic method of the differential privacy technology is used in this process. The specific realization process is as follows: Calculate the gradient value $g(x)$ of each random sample; In order to avoid the impact of a single data on the whole, the gradient was adjusted, and the norm L_2 of each gradient was clipped and the average value of the gradient was calculated to meet the threshold condition C, so as to obtain a new gradient value $\overline{g}$. In order to realize privacy protection, noise is added to the new gradient value $\overline{g}$ to interfere with the output of the gradient. Finally, the gradient with added noise is increased in the opposite direction according to the gradient descent method, the parameter θ is updated, and the privacy loss is calculated.

Algorithm 2.1 Depth learning differential privacy algorithm

Input: training sample set $\{x_1, x_2, \ldots, x_n\}$, Loss function $L(\theta)=\frac{1}{N}\sum_i L(\theta, x_i)$. To add the size of the noise σ, learning rate α, the gradient clipping threshold C, the number of times the training was conducted T, the size of the random sample group Initialization m.

random value θ_0

for $t\in[T]$ do

Random extraction m_t training sample

for $x_i \in m_t$ do

$g_t(x_i) \leftarrow \nabla\theta_t L(\theta_t, x_i)$

$$\overline{g_t} \leftarrow \frac{g_t(x_i)}{\max\left[1, \frac{\| g_t(x_i) \|}{C}\right]}$$

$\tilde{g_t} \leftarrow \overline{g_t} + N$

$\theta_{t+1} \leftarrow \theta_t - \eta\tilde{g_t}$ end for

Output: Gradient value θ_t, Using the statistical method of differential privacy to calculate the privacy loss.

2.2 privacy loss analysis

In order to evaluate the privacy protection performance of the deep learning differential privacy protection model reasonably, we need to make a statistical analysis of the privacy loss in the training process. In this process, parameters are usually updated many times due to the decline of the gradient, which will lead to a gradual accumulation of the privacy loss. Combining the characteristics of differential privacy, this paper uses the time accounting method[6] proposed by Abadi and others for a calculation of the privacy loss. Privacy loss is regarded as a random variable whose value depends on the noise added to the algorithm. By calculating the logarithmic moment of the random variable Z of privacy loss, and using the time limit and the standard Markov inequality to obtain the tail limit, we can get the privacy loss. Then the privacy loss of the random variable Z is defined as

$$Z(s, M, aux, D, D') \triangleq \log \frac{\Pr[M(D)=s]}{\Pr[M(D')=s]} \quad (5)$$

Among them, $M$is a random algorithm , D and D' are the nearest neighbor data , aux is used to assist the input, and s is the output.

2.3 The advantages of WGAN

There are some problems in the original GAN, such as the instability of training and the lack of generating diversity due to the mode collapse. GAN Trains G and D in an alternating optimization mode, and the optimization between G and D must achieve a good synchronization. However, in the actual training process, usually after D is updated many times, G will be updated once, which is easy to cause the G collapse to a saddle point. Some scholars try to solve the problem with DCGAN. Although DCGAN has a good structure, if the G of DCGAN loses the batch standardization, it will lead to the collapse of the training. The problem of instability of GAN Training has not been solved completely. Arjovsky et al[13] proposed Wasserstein GAN to improve on the shortcomings of the original GAN. WGAN uses Wassertein distance as an optimization method instead of the cross entropy to measure the distance between the real distribution and the generated distribution, so that the convergence tends to be stable and the training stability is greatly improved. Compared with DCGAN, WGAN is not restricted by batch standardization, and can use special network to achieve G and D, thus getting more diverse generation effects. Therefore, WGAN is chosen to replace the original GAN, and appropriate noise is added to the gradient in the process of deep learning training to realize privacy protection under WGAN.

2.4 The privacy feedback based on WGAN

Privacy feedback is defined as a way to effectively establish contact with model-related parameters through feedback channels and obtain timely responses. Privacy feedback reflects a two-way flow of information. The deeper the feedback, the more favorable it is to choose the appropriate privacy parameter settings.

The realization process of the feedback mechanism in this paper is to judge the threshold by comparing the similarity between the generated result G_1 and the generated result G_2, and adjust the relevant parameters of the depth differential privacy by combining the evaluation criteria of privacy loss and accuracy. Without privacy feedback, the privacy parameters in depth learning differential privacy will be uncontrollable, and the generated results will have the defects of insufficient privacy protection or greatly reduced data availability. The basic framework is shown in Fig. 1.

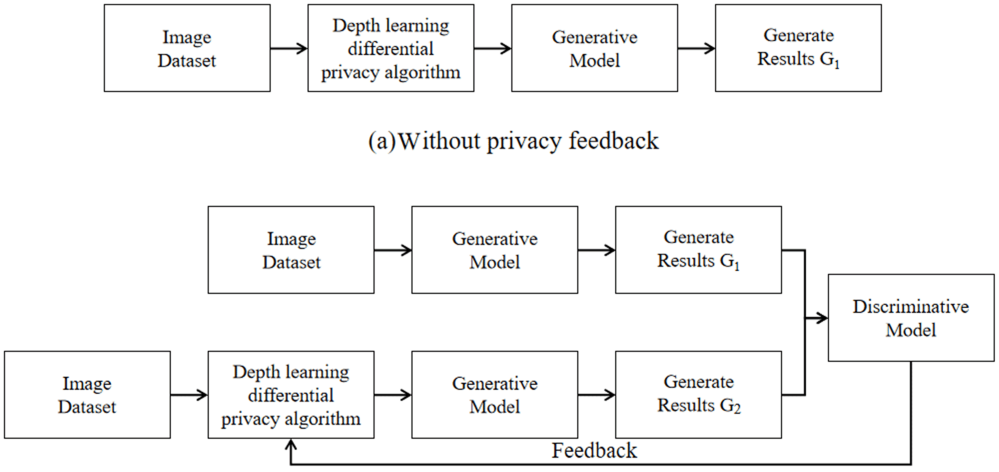

Fig. 1 Flow chart of the deep learning differential privacy

3 Experiment and analysis

3.1 Datasets

In this paper, the MNIST handwritten digital dataset[14] is selected as the experimental dataset. The MNIST dataset is a handwritten digital database established by Corinna Cortes of Google Laboratory and Yann LeCun of Kiran Institute of New York University. The training database contains 60000 handwritten digital pictures and corresponding tags, and the test database has 10000. Each picture consists of 28×28 pixels.

3.2 Experiments

In this experiment, a random gradient drop is used to be optimized, and the gradient threshold C is set at 5 in the iterative process. Using the MNIST training set under different differential privacy parameters set, using WGAN to train and generate model and generate data, and then train the generated data to get classifier and calculate the test accuracy for MNIST test set. At the same time, the classifier is trained for the original MNIST dataset without differential privacy processing, and the test set is used to calculate the test accuracy. Compare the difference between the image generated by differential privacy protection WGAN and the unprocessed image, use the moment accounting method to determine the loss of privacy, on the premise of ensuring a certain degree of privacy protection, select the appropriate parameters to improve the test accuracy of the model.

3.2.1 Changes ε influence on the experiment

The definition of differential privacy indicates the privacy budget. The smaller the value is, the better the degree of privacy protection. In order to verify the effect of the change of the ε value on the accuracy of the test, the experiment fixed $\delta = 10^{-5}$ and $\sigma = 6$, the privacy budget parameters ε were changed from 0.5 to 8. The experimental results are shown in the following figure. When the privacy budget parameters ε are 0.5, 2, 4 and 8, the experimental accuracy is 89.43%, 90.19%, 90.86%, 93.05% respectively. The results shown in Fig. 2 which shows that the accuracy of the test increases with the increase of privacy budget parameters ε. In addition, the selection of the privacy budget parameters ε should not be too high, or it will add too little noise and affect the effect of privacy protection.

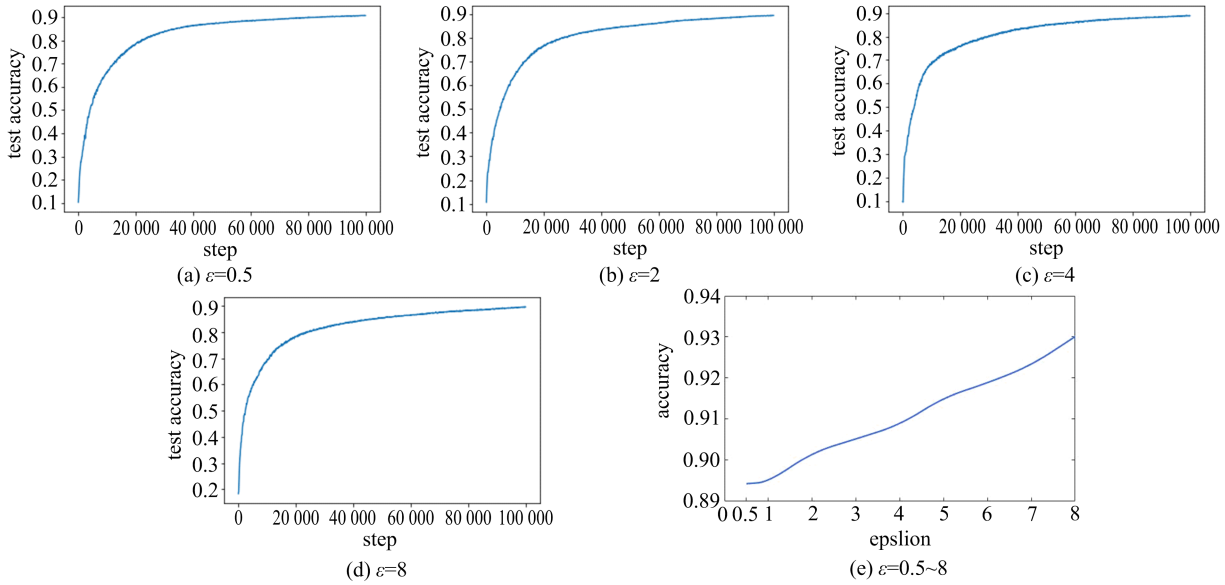

Fig. 2　Influence of test accuracy by changing single variable ε

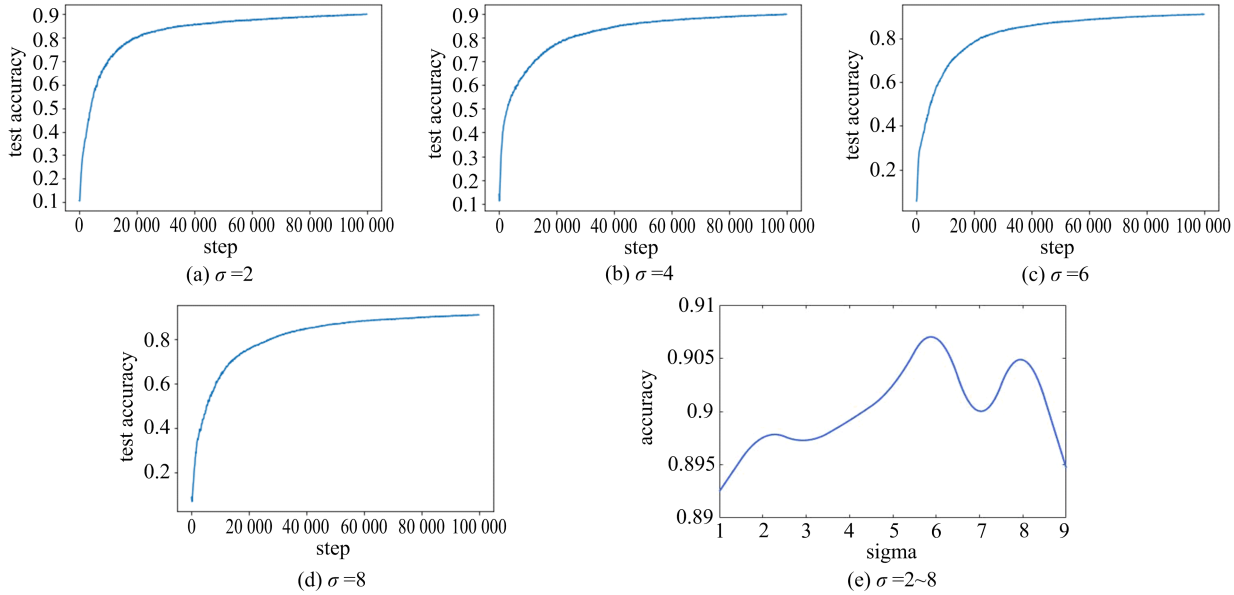

Fig. 3　Influence of test accuracy by changing single variable σ

3.2.2　Changes σ influence on the experiment

The second group of experiments fixed$\varepsilon=0.5$ and $\delta=10^{-5}$ to verify the effect of changing the noise scale σ on the accuracy of experimental tests. The acceptable values of the noise scale σ are 1 to 9 respectively. The experimental results are shown in the Fig. 3. When σ is taken as 2, 4, 6 and 8 respectively, the test accuracy of the experiment is 89.85%, 89.91%, 90.93% and 90.22% respectively.

The results shown in Fig. 3 which shows that the increase of the σ value has an alternating trend of increasing first and then decreasing on the accuracy of the model. When σ value is 6, the accuracy of the model test is relatively high.

3.2.3　Changes δ influence on the experiment

The third group of experiments fixed $\varepsilon=0.5$ and $\sigma=6$ to verify the effect of changing the privacy leak error δ on the accuracy of experimental tests. The experimental results are shown in the following figure. When δ is taken as 10^{-2}, 10^{-3}, 10^{-4} and 10^{-5} respectively, the test accuracy of the experiment is 89.41%, 89.71%, 89.89% and 90.19% respectively.

The results shown in Fig. 4 which shows that, with the decrease of privacy leak error δ, the test accuracy of the model is slightly improved. When the value of δ is 10^{-5}, there is a balance between the privacy leak error and the model test accuracy, and the test accuracy is relatively high at this time.

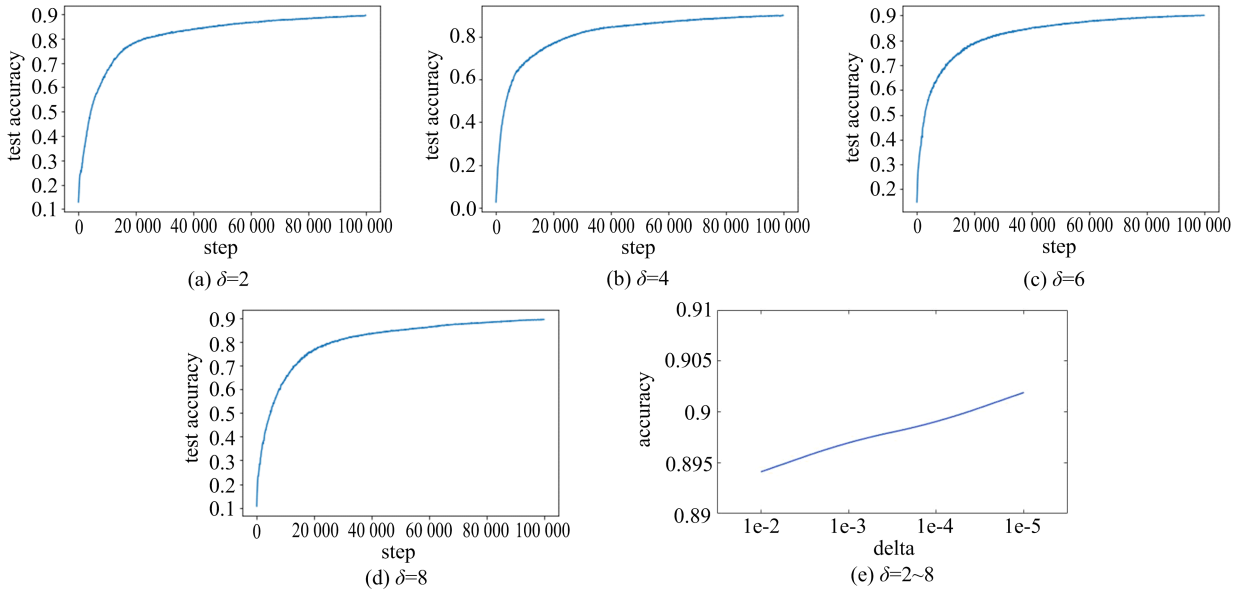

Fig. 4 Influence of test accuracy by changing single variable δ

3.2.4 Changes σ and δ influence on the experiment

In the fourth group of the experiment, the $\varepsilon = 0.5$ was fixed and the values of σ and δ were changed, in which the values of σ were 2, 4, 6 and 8 respectively, and the values of δ were 10^{-5} to 10^{-2}. The experimental results are shown in Fig. 5.

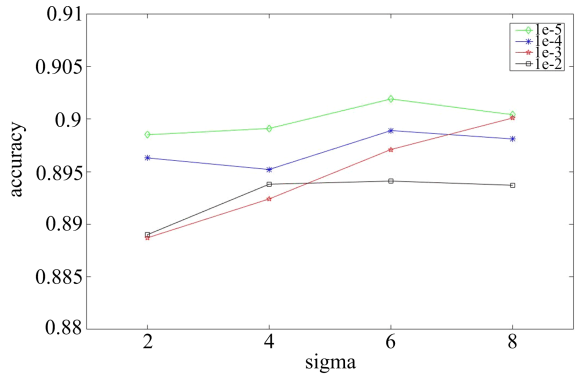

Fig. 5 Influence of test accuracy by changing single variable σ and δ

3.2.5 Changes σ and ε influence on the experiment

In the fifth group of experiment, the $\delta = 10^{-5}$ was fixed and the values of σ and ε were changed, in which the values of σ were 2, 4, 6 and 8 respectively, and the values of ε were 0.5, 1, 2 and 4 respectively. The experimental results are shown in Fig. 6.

3.2.6 Changes δ and ε influence on the experiment

In the sixth group of the experiment, the $\sigma = 6$ was fixed and the values of δ and ε were changed, in which the values of δ were 10^{-5} to 10^{-2}, and the values of ε were 0.5, 1, 2 and 4 respectively. The experimental results are shown in Figs. 7 and 8.

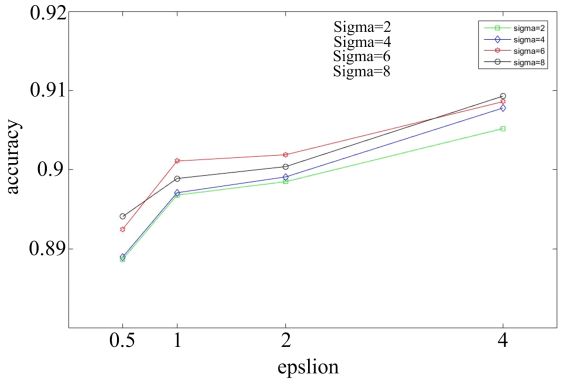

Fig. 6 Influence of test accuracy by changing single variable σ and ε

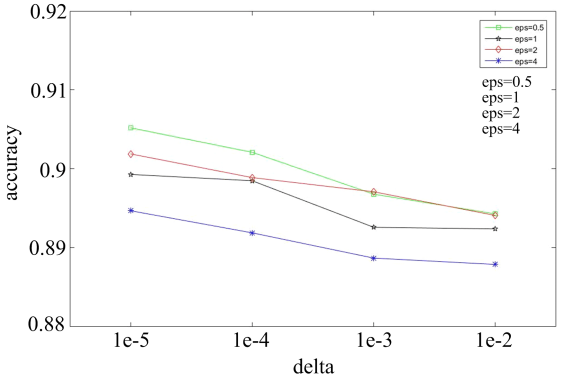

Fig. 7 Influence of test accuracy by changing single variable δ and ε

4 Conclusion

Through the experiments of several sets of control parameter variables, we verified the separate influence of the change of the single related parameter on the experimental test accuracy and the mutual influence of the simultaneous change of different related

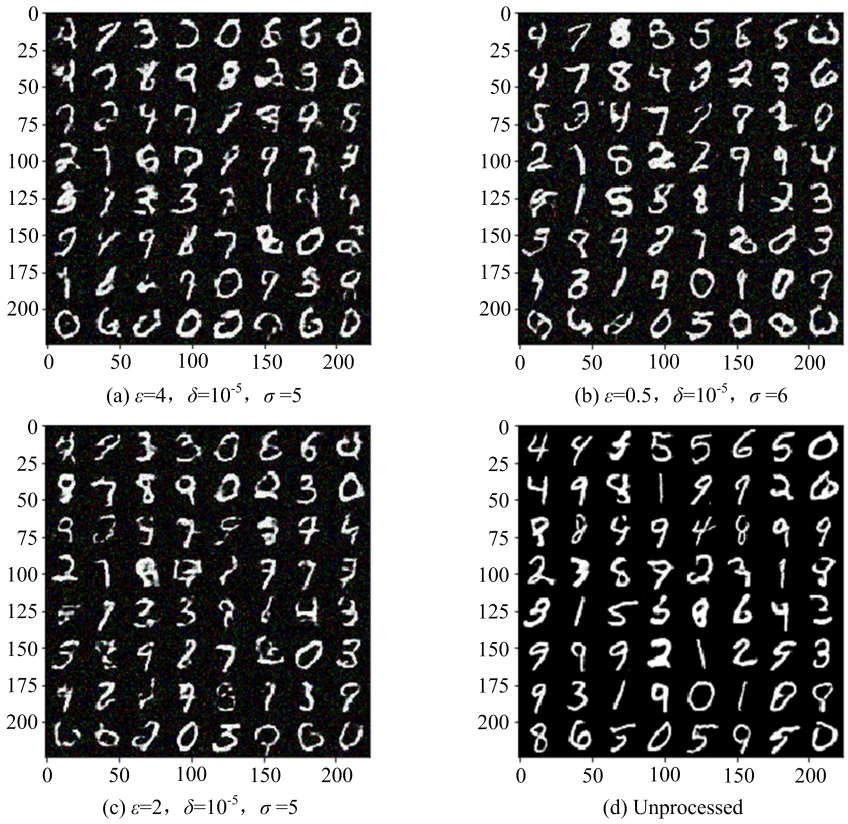

Fig. 8　Generated images corresponding to different privacy parameters

parameters on the experimental test accuracy. It was found that when ε takes the value of 0.5, δ takes the value of 10^{-5}, σ takes the value of 6, the experimental test accuracy of the deep learning differential privacy protection model is 90.52%, and the privacy budget calculated by the time accounting method can ensure the protection of privacy. Compared with other GAN, the training process using WGAN is more stable and the accuracy rate is improved to a certain extent, which basically realizes the balance between the degree of privacy protection and the availability of datasets.

References

[1] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7):1527-1554.

[2] SWEENEY L. K-anonymity: A model for protecting privacy[J]. International Journal of Uncertainty: Fuzziness and Knowledge-Based Systems, 2002, 10(05):557-570.

[3] MACHANAVAJJHALA A, KIFER D, GEHRKE J. L-diversity: Privacy beyond k -anonymity[J]. Acm Transactions on Knowledge Discovery from Data, 2007, 1(1): 3.

[4] LI N, LI T, VENKATASUBRAMANIAN S. t-Closeness: Privacy Beyond *k*-Anonymity and l-Diversity[C]//IEEE 23rd International Conference on Data Engineering. Piscataway: IEEE, 2007.

[5] DWORK C. Differential privacy[J]. Lecture Notes in Computer Science, 2006, 26(2):1-12.

[6] ABADI M, GOODFELLOW I, GOODFELLOW I, et al. Deep Learning with Differential Privacy[C]// ACM Sigsac Conference on Computer & Communications Security. 2016.

[7] PAPERNOT N, ABADI M, ERLINGSSON L, et al. Semi-supervised knowledge transfer for deep learning from private training data[C/OL]. (2017-03-03)[2020-05-05]. https://arxiv.org/pdf/1610.05755v4.

[8] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// International Conference on Neural Information Processing Systems. 2014.

[9] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. Computer Science, 2015.

[10] ULLOA A, BASILE A, WEHNER G J, et al. An unsupervised homogenization pipeline for clustering similar patients using electronic health record data[C/OL]. (2018-03-21)[2020-05-05]. https://arxiv.org/pdf/1801.00065.

[11] XIE L, LIN K, WANG S, et al. Differentially private generative adversarial network[C/OL]. (2018-02-19)[2020-05-05]. https://arxiv.org/pdf/1802.06739.

[12] HITAJ B, ATENIESE G, PEREZ-CRUZ F. Deep models under the GAN: Information leakage from collaborative deep learning[C]// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: Association for Computing Machinery, 2017: 603-618.

[13] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[C/OL]. (2017-12-06)[2020-05-05]. https://arxiv.org/pdf/1701.07875.

[14] LECUN Y, CORTES C, BURGES C J. The MNIST database of handwritten digits[DB/OL]. [2020-05-05]. http://yann.lecun.com/exdb/mnist/? o=3510.