IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 July 2023, accepted 12 September 2023, date of publication 14 September 2023, date of current version 6 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3315592

RESEARCH ARTICLE

Differentially Private Denoise Diffusion Probability Models

ZHIGUANG CHU[1,2], JINGSHA HE[1], DONGDONG PENG[2], XING ZHANG[2], AND NAFEI ZHU[1]

[1]Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[2]Key Laboratory of Security for Network and Data in Industrial Internet of Liaoning Province, Jinzhou 121000, China

Corresponding author: Nafei Zhu (219915010@stu.lnut.edu.cn)

This work was supported in part by the Applied Basic Research Project of Liaoning Province under Grant 2022JH2/101300280, and in part by the Scientific Research Project of Liaoning Provincial Department of Education under Grant LJKZ0625.

ABSTRACT Diffusion models and their variants have achieved high-quality image generation without adversarial training. These algorithms provide new ideas for data shortages in some fields. But the diffusion model also faces the same problem as other generative models: the learned probability density function will retain the characteristics of the training samples, which means that the high complexity of the deep network will make the model easily remember the training samples. When a diffusion model is applied to sensitive datasets, the distribution the model focuses on may reveal private information, and the security concerns described above become more pronounced. To address this challenge, this paper proposes a privacy diffusion model named DPDM (Differentially Private Denoise Diffusion Probability Models) that satisfies differential privacy by adding appropriate noise to the gradient during the training. Besides, this paper adopts a series of optimization strategies to improve model performance and training speed such as adaptive gradient clipping threshold and dynamic decay learning rate. Through the evaluation and analysis of the benchmark dataset, it is found that the attempt in this paper has promising usability, and the synthetic data has better performance.

INDEX TERMS Data shortage, generate model, diffusion model, differential privacy.

I. INTRODUCTION

The ever-increasing data scale and the continuous innovation of Internet technology promote each other. As a representative of the latter, deep learning also faces more problems such as data shortages in some application fields. For example, in the analysis of individual patients in the medical field, each patient can be regarded as an individual sample in the model training process. Due to the diversity and complexity of diseases, there are only a handful of patient records after refinement, which is difficult to use as the basis for research. Furthermore, data holders are reluctant to share data due to privacy concerns, which will exacerbate the problem of research data scarcity. When the data is used to train the deep model, some sensitive features are also more likely to be remembered. Individual examples have proved that the memory of the model in deep learning can effectively restore the sensitive information of the training data [1], which will cause the deep model to become the target of attacks by people with ulterior motives.

As the gold standard in the field of machine learning, differential privacy [2], [3], [4] can quantify privacy strictly and provides an important idea for solving privacy problems. In 2016, the DP-SGD algorithm proposed by Abadi et al. [5] immediately became an idiomatic method for training differential privacy learners. The core step of DP-SGD (Differentially Private Stochastic Gradient Descent) is to clip the gradient norm and inject Gaussian noise. Nowadays, more and more researchers apply the DP-SGD algorithm to the training process of deep learning models [6], [7], [8], [9], [10], [11], [12]. After differential privacy training, even if the attacker obtains the model, one cannot infer the private information of the training set.

Generative adversarial networks (GANs) and variants [13], [14], [15] can generate more fake samples that are indistinguishable from the training data. But there are still

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

shortcomings: its training is still not stable enough, and it needs to train two network models at the same time, that is, the adversarial training between the generator and the discriminator, which will inevitably lead to balance problems and easily lead to model collapse. Due to GAN's pursuit of image authenticity, at the same time, the input of the generator is random noise, resulting in insufficient diversity.

Diffusion models [16], [17], [18], as one of the emerging technologies in the field of generation, gradually interfere with the data in the forward process, while the neural network predicts the noise added during the forward process and gradually removes to learn the training distribution. Compared with GAN, the diffusion model is not only more stable in training; but also achieves better diversity.

In order to deal with the training and diversity problems of GAN, as well as the privacy leakage problem of the diffusion model, this paper combines the differential privacy [2], [3] and the diffusion model [16], [17], [18] to design a privacy protection framework DPDM to protect the model data. In summary, the contributions of this paper can be summarized in the following three points:

- This paper proposes a DPDM privacy-preserving framework by combining differential privacy and diffusion models, which provide privacy protection for models while ensuring the availability of synthetic data.
- During the training, a series of optimization strategies such as gradient threshold dynamic clipping and learning rate degradation algorithms, are adopted to speed up the convergence and improve the quality of synthetic data onto a given privacy budget.
- Our method in this paper yields state-of-the-art performance on model utility under the same privacy budget. By using the benchmark dataset to evaluate the model and verify the performance of DPDM, it is proved that DPDM can achieve better diversity and visual effects.

The remainder of this paper is organized as follows. In Section II is the related work, and we describe the preliminaries in Section III. In Section IV we present our approach in detail, while Section V shows our experimental results. Conclusion of this paper in Section VI.

II. RELATED WORKS

Many researchers have tried to protect training data. Prior techniques include data anonymization [19], k-anonymity [20], l-diversity [21], t-closeness [22], semantic security [23], information-theoretic privacy [24], and differential privacy (DP) [2], [3], [4], where DP is a rigorous mathematical definition of privacy applied to statistical queries; in our work the queries correspond to the training of a neural network using sensitive training data. Once the DP-SGD [5] algorithm was proposed, it became the mainstream of machine learning private training. The main steps of the DP-SGD algorithm are gradient norm clipping and Gaussian noise injection into the gradient. The differential privacy learner aims to learn the same distribution as the private data while satisfying the differential privacy guarantee.

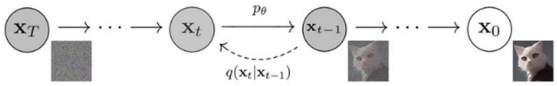

FIGURE 1. Illustration of diffusion models.

A long line of works has explored combinations of GANs and differential privacy [6], [7], [8], [9], [10], the injection of noise makes GAN require a lot of optimization strategies and structural design to perform more stable, and the generation effect is not satisfactory. (Differential Privately Generate Adversarial Nets) DPGAN [6] first combines differential privacy and GAN by DPSGD algorithm, where the discriminator is privately trained, and the generator automatically satisfies differential privacy by the post-processing theory. Torkzadehmahani et al. [7] proposed to use DPSGD to train conditional GAN to extend DPGAN to the conditional generative setting. Zhang et al. [8] used differential privacy in GAN. When training the discriminative model, the DP-SGD algorithm is used to disturb the gradient. The game between the generator and the discriminator is able to improve the quality of the generation. Xu et al. [9] proposed a GAN-obfuscator to promote, by designing a gradient penalty strategy to achieve high-quality synthetic data generation. Private Aggregation of Teacher Ensembles (PATE) is a different framework for generative models. Reference [10] proposed PATE-GAN, adapting the PATE framework to apply it to GAN to generate data. G-PATE [11] ensures that the information flow from the discriminator to the generator is private so that achieves differential privacy guarantee. Although the above methods improve and perfect the differential privacy protection method of generative models, the problem of private training difficulty and the low availability of synthetic data has not yet been solved.

III. PRELIMINARIES

A. DIFFUSION MODELS

We continue our research based on Improved Denoising Diffusion Probabilistic Models (IDDPM) [17] and use its probabilistic representation.

As shown in Fig. 1, its forward process adds noise to the input data until the signal is destroyed, resulting in noise that follows normal distribution. Its reverse process samples from the normal distribution to obtain noise and then continuously denoises it until it is restored to data that follows the target distribution.

Definitions: Given a data distribution $x_0 \sim q(x_0)$, we can obtain latent variable x_1 to x_T through a forward process q represented by Gaussian transitions at time t:

$$q\left(x_1, \ldots, x_T \mid x_0\right) = \prod_{t=1}^{T} q\left(x_t | x_{t-1}\right) \tag{1}$$

$$q\left(x_t \mid x_{t-1}\right) = N\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t \boldsymbol{I}\right) \tag{2}$$

where variance schedule $\beta_t \in (0, 1)$. According to Equation 2, conditioned on the input x_0, we can sample an

arbitrary step of noised latent variables. With $\alpha_t := 1-\beta_t$, $\bar{\alpha}_t := \prod_{i=0}^{t} \alpha_i$, the marginal can be written as follow:

$$q(x_t \mid x_0) = N\left(x_t; \sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t)\boldsymbol{I}\right) \tag{3}$$

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{4}$$

where ϵ is a random noise that follows normal distribution, $\epsilon \sim N(0,\boldsymbol{I})$. $1-\bar{\alpha}_t$ denoted the variance of the noise for an arbitrary step and can be used to defined the noise schedule instead of β_t.

The reverse process uses a neural network to predict the reverse distribution as follows:

$$p_\theta(x_{t-1} \mid x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \sigma(x_t, t)) \tag{5}$$

where reverse process p can also be represented by Gaussian transitions. By predicting the mean and variance through neural network, an approximation of the distribution of the previous step is estimated. This process is repeated iteratively until the distribution of the training data is obtained.

Training: The combination of the forward process and the reverse process is a variational auto-encoder L_{vlb}, which is consists of KL divergence between two Gaussians [25]. Besides, Ho et al. [26] found that predicting worked best when combined with a reweighted loss function L_{simple}. Using Bayes theorem, the posterior $q(x_{t-1} \mid x_t, x_0)$ can be calculated. We can use the prior (Equation 5) and the posterior (Equation 9) to estimate L_{vlb}. Therefore, this paper adopts a hybrid objective function to optimize neural network:

$$L_{hybrid} := L_{simple} + \lambda L_{vlb} \tag{6}$$

where:

$$\begin{aligned} &L_{vlb} \\ &:= \mathbb{E}_q \Big[D_{KL}(q(x_T \mid x_0) \| p_\theta(x_T)) + \sum_{t=2}^{T} \\ &\quad \times D_{KL}(q(x_{t-1} \mid x_t, x_0) \| p_\theta(x_{t-1} \mid x_t)) - \log p_\theta(x_0 \mid x_1) \Big] \end{aligned} \tag{7}$$

$$L_{simple} := \mathbb{E}_{t, X_0, \varepsilon}\left[\left\| \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right) \right\|^2 \right] \tag{8}$$

$$\begin{aligned} &q(x_{t-1} \mid x_t, x_0) \\ &= N\left(x_{t-1}; \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t, \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \boldsymbol{I}\right) \end{aligned} \tag{9}$$

B. DIFFERENTIAL PRIVACY

Differential privacy [2], [3], [4] is a privacy protection technology based on data disturbance. It interferes with private data by injecting noise, while ensures that the utility of the published data. Differential privacy is defined as follows:

Definition 1 (Differential Privacy [2], [3], [4]): A randomized mechanism $M{:}D \to R$ with domain D and range R satisfies $(\varepsilon, \delta)-$ Differential Privacy if for any datasets $d, d' \in D$ differing by at most one entry, and for any subset of outputs holds that:

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

where ε is privacy constrict, which indicates the degree of privacy protection. δ indicates the probability of privacy leakage under differential privacy. The smaller is ε, the greater the degree of privacy protection is, and the corresponding usability is lower.

Definition 2 (RényiDifferentialPrivacy (RDP) [27]): A randomized mechanism M is (λ, ε)-RDP with the order λ, if:

$$\begin{aligned} &D_\lambda\left(M(d) \| M(d')\right) \\ &\quad = \frac{1}{\lambda-1} log\, \mathbb{E}_{x \sim M(d)}\left[\left(\frac{P[M(d)=x]}{P[M(d')=x]}\right)^{\lambda-1}\right] \leq \varepsilon \end{aligned}$$

holds for any adjacent dataset d and d', where $D_\lambda(P \| Q) = \frac{1}{\lambda-1} log E_{x \sim Q}\left[\left(\frac{P(x)}{Q(x)}\right)^{\lambda}\right]$ denotes the *Rényi* divergence, P and Q are the probability density function. Moreover, a $(\lambda, \varepsilon) - RDP$ mechanism M also satisfies $\left(\varepsilon + \frac{\log\left(\frac{1}{\delta}\right)}{\lambda-1}, \delta\right) - DP$.

Theorem 1 (Composition [27]): For a sequence of mechanisms $M_1, \ldots, M_k$ s.t. M_i is $(\lambda, \varepsilon) - RDP \forall i$, the composition $M_1 \circ \ldots \circ M_k$ is $\left(\lambda, \sum_i \varepsilon_i\right) - RDP$.

Definition 3 (Gaussian Mechanism [2], [3], [4]): Let $f : X \to R^d$ be an arbitrary d-dimensional function with sensitivity being:

$$\Delta f = max_{d,d'} \left\| f(d) - f(d') \right\|$$

over all adjacent datasets d and d'. The Gaussian Mechanism M_σ, parameterized by σ, adds noise into the output, i.e,

$$M_\sigma(d) = f(d) + N(0, \sigma^2 \Delta f^2 \boldsymbol{I})$$

M_σ is $\left(\lambda, \frac{\lambda \Delta_2 f^2}{2\sigma^2}\right) - RDP$.

IV. DESIGN OF FRAMEWORK

In this section, we introduce the DPDM framework, which combines state-of-the-art learning techniques with advanced privacy-preserving mechanisms. The DPDM framework, shown in Fig. 2, aims to address privacy challenges while maintaining usability. The curator uses a privacy-preserving method to train the diffusion model, in which Abadi et al. [5] 's DPSGD is used to add noise to the gradient during training, trained model satisfies differential privacy. Instead of releasing a sanitized version of the dataset, the curator releases a differentially private generative model. Equipped with generative model, analyst can generate unlimited synthetic data for their intended analysis tasks. The utilization of differential privacy guarantee that no one can deduce the original training data from the published generative model, thus ensuring the privacy protection of the whole process.

We achieve private training when predicting noise in the reverse process. Similar to the work of [6], when executing the gradient descent algorithm, the gradient is clipped and added noise. This paper focuses on preserving privacy during the training phase rather than adding noise directly to

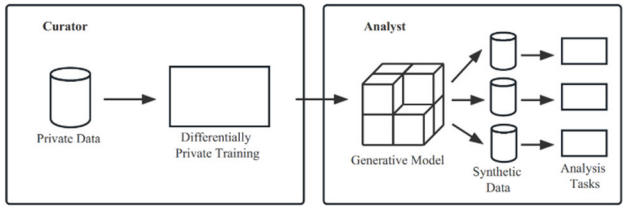

FIGURE 2. DPDM privacy protection publishing framework.

Algorithm 1 Training

Input: $\{x_i\}_{i=1}^m \sim p_{data}$: a batch of training samples; σ_{DP}: DP noise scale; m: batch size; $(\alpha, \beta_1, \beta_2)$: Adam hyper-parameters; (ϵ, δ): overall privacy target; θ: initial model; T: total steps;d: decay rate;

Output: model θ

1: **for** $t = 1$ **to** T **do**
2: Compute Gradient
3: $g_{batch} \leftarrow \nabla_\theta L\left(\theta, \{x_i\}_{i=1}^m\right)$
4: Clip Gradient
5: $C \leftarrow$ *Adaptive Threshold* $\left(\{x_i\}_{i=1}^m\right)$
6: $\bar{g}_{batch} \leftarrow g_{batch}/max\left(1, \frac{\|g_{batch}\|_2}{C}\right)$
7: Add Noise
8: $\tilde{g}_{batch} \leftarrow \left(\left(\frac{1}{m}\right)\sum_m \bar{g}_{batch} + (C/m)N\left(0, \sigma_{DP}^2\right)\right)$
9: Calculate Privacy Loss
10: $\alpha \leftarrow$ *Learning Rate Degrading* (t, d)
11: $\theta \leftarrow Adam\ (\theta, \alpha, \beta_1, \beta_2)$
12: **end for**
14: **return** model θ

the final parameters. However, unlike [6], we optimize the clipping process by selecting a dynamic gradient threshold. Compared with existing methods, our model converges faster and generates images with better availability and diversity.

Algorithm 1 gives the key steps in DPDM.

As shown in Algorithm 1, during the model optimization, we first sample a batch of samples from the training dataset. We compute the model gradients based on this batch of samples (line 2-3). Before adding noise to the gradients, we use an adaptive clipping threshold algorithm to obtain a clipping threshold to get tighter privacy loss, which is used to clip the gradients to control the sensitivity (line 4-6). After clipping, Gaussian noise is added to the gradients (line 7-8). The learning rate obtained by a learning rate decay algorithm and optimizer parameters are passed to the optimizer (line 10-11). Finally, iterations are performed, and the model parameters are updated and optimized. Besides, in order to achieve the target level of privacy protection, we track the cumulative privacy loss, if privacy loss exceeds the privacy budget, the training is terminated.

A. LEARNING RATE DECAY

During the training of a deep learning model, the learning rate is a hyperparameter that controls the adjustment speed of the neural network weights based on the gradient of the loss. The choice of learning rate has a trade-off between the accuracy of the model and the speed of convergence: if the value is larger, the training speed will increase but the model accuracy will be insufficient; correspondingly, if the learning rate is smaller, although the accuracy will increase, the convergence of the model leads to higher time complexity. Therefore, in order to enable the algorithm to achieve better performance, in the early stage of training, this paper sets a larger learning rate to make the loss function approach the optimal value as soon as possible. After training for a while, use a smaller learning rate to improve the model accuracy. We tried several learning rate decay strategies as follows:

$$\text{Linear decay}: \ lr = \frac{lr'}{1 + dr * epoch};$$
$$\text{Exponential decay}: \ lr = lr_{initial} * dr^{epoch};$$
$$\text{Logarithmic decay}: \ lr = \frac{lr'}{1 + log(epoch)};$$
$$\text{No decay}: \ lr = lr_{fixed};$$

In the above functions, lr indicates the current learning rate; lr' indicates the last learning rate; dr is the decay rate; $epoch$ is the current epoch; $lr_{initial}$ is the initial learning rate;lr_{fixed} is the fixed learning rate. In our experiment, we show different results for different strategies.

B. ADAPTIVE CLIPPING THRESHOLD

In the gradient clipping of DPSGD algorithm, if the L2 norm of the sample gradient is more than the predetermined clipping threshold C, the L2 norm of the gradient gets scaled down to be of norm C; if it is smaller than C, the gradient is preserved. By gradient clipping, we can obtain the sensitivity of the gradient aggregate with respect to the addition or removal of any sample, adding Gaussian noise to the aggregated gradients, and thus achieving differential privacy guarantee. And the selection of the clipping threshold has a great impact on the performance of the optimizer: if C is too small, it will lead to excessive clipping of the gradient, resulting in slow convergence and even model training failure; if C is too large, too much noise will be added to the gradients leading to poor quality of the final generated image. And because the weights and biases of different network layers are quite different, the gradient also changes with the training, and it is difficult to find the optimal clipping threshold. To remedy such issues, this paper introduces a dynamic strategy, which can be used to adjust the clipping threshold. Zhang et al. [8] proposed an adaptive gradient clipping method. They clustered parameters such as weight and bias, monitored the changes in gradient values before and during training, and then set gradient thresholds for each cluster. Based on the Diffusion Model framework, this paper follows the ideas of Zhang et al. [8], assuming that besides private data D_{pri} to train the model, our algorithm has access to a small amount of public data set D_{pub}. During each training of step, we randomly sample a batch of samples from D_{pub},

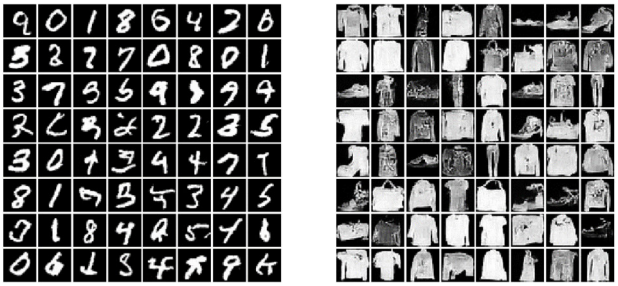

FIGURE 3. DPDM synthetic data on MNIST and Fashion MNIST.

set the average gradient norm of this batch as the gradient clipping threshold.

C. PRIVACY ANALYSIS

Our privacy computation is based on the notion of *Rényi* Differential Privacy. To obtain the overall accumulated privacy cost over multiple training iterations, we use the composition properties of RDP. In this section, we provide a short proof that the gradients released by the Gaussian mechanism are DP. For noise scale σ_{DP}, one batch of data $\{x_i\}_{i\in\mathbb{B}}$ and $\{x_i\}_{i\in\mathbb{B}}\cup x', x' \notin \{x_i\}_{i\in\mathbb{B}}$. We can bound the difference of their gradients in $L_2 - norm$ as:

$$\begin{aligned}
&\left\|g_{batch}\left(\{x_i\}_{i\in\mathbb{B}}\right) - g_{batch}\left(x'\cup\{x_i\}_{i\in\mathbb{B}}\right)\right\|_2 \\
&= \Big\|1/m\sum\nolimits_{i\in\mathbb{B}} clip_C\left(\nabla_\theta L\left(x_i\right)\right) - (1/m clip_C\left(\nabla_\theta L\left(x'\right)\right) \\
&\quad +1/m\sum\nolimits_{i\in\mathbb{B}} clip_C\left(\nabla_\theta L\left(x_i\right)\right)\Big\|_2 \\
&= \left\|-1/m clip_C(\nabla_\theta L\left(x'\right))\right\|_2 \\
&= 1/m\left\|clip_C(\nabla_\theta L\left(x'\right))\right\|_2 \le C/m
\end{aligned}$$

where m denotes the batch size, $clip_C$ denotes the clipped gradient. We thus have sensitivity C/m. Furthermore, since $z\sim N(0,\sigma_{DP}^2)$, $(C/m)z\sim N(0,(C/m)^2\sigma_{DP}^2)$. Following standard arguments, releasing $\tilde{g}_{batch}\left(\{x_i\}_{i\in\mathbb{B}}\right) = g_{batch}\left(\{x_i\}_{i\in\mathbb{B}}\right)+(C/m)z$ satisfies $(\alpha, \alpha/2\sigma_{DP}^2)-$ RDP.

V. EXPERIMENTAL EVALUTION

In this section, this paper evaluates DPDM privacy-preserving framework experimentally. First, this paper verifies whether DPDM can synthesize visual realistic images under differential privacy. Secondly, this paper determines whether it can generate high-quality and diverse image datasets. Finally, this paper applies synthetic images to perform classification tasks to verify the practicability.

A. DATASETS

The experiment is carried out on the benchmark datasets (MNIST [28], Fashion-MNIST [29], CelebA [30]). For MNIST and Fashion-MNIST, both are consist of 60,000 training samples and 10,000 test samples. Each sample is a 28×28 grayscale image, divided into 10 categories. CelebA is a dataset including face images of celebrities. Each image is a 178×218 RGB image, and has 40 binary attributes. These

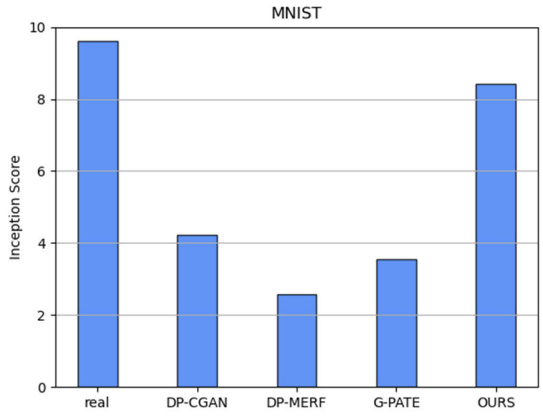

FIGURE 4. Inception Score of real images and synthetic images on MNIST.

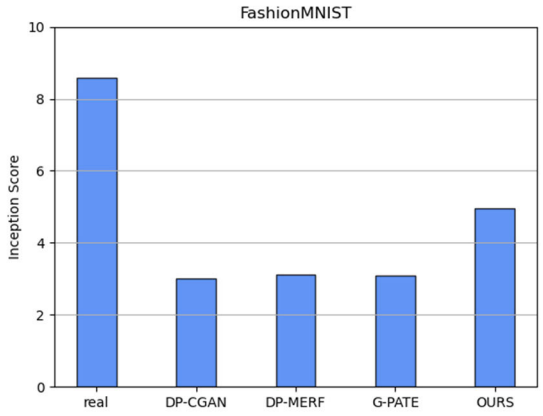

FIGURE 5. Inception Score of real images and synthetic images on Fashion MNIST.

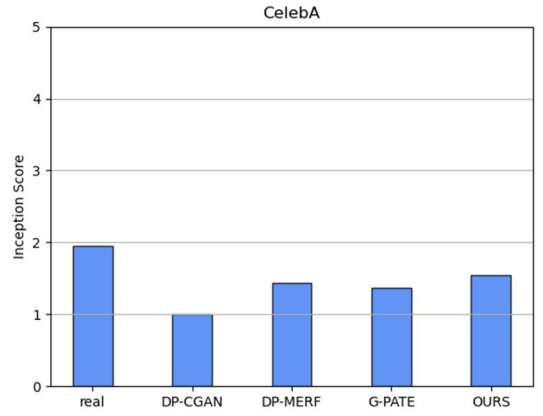

FIGURE 6. Inception Score of real images and synthetic images on CelebA.

datasets have become the most commonly used benchmark dataset in deep learning model research.

B. METRICS

In all experiments, we consider Inception Score (IS) [31], [32], which is used to evaluate the quality of generated images by generative models. It combines the diversity and the realism of the generated images to quantifies the performance

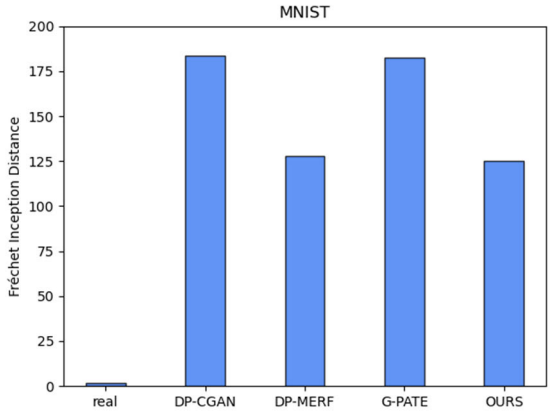

FIGURE 7. Fréchet Inception Distance of real and synthetic data on MNIST.

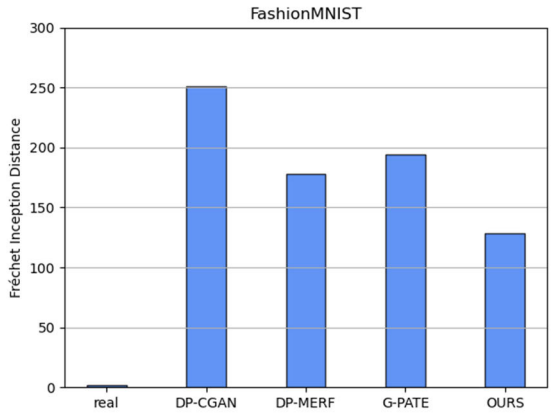

FIGURE 8. Fréchet Inception Distance of real and synthetic data on Fashion MNIST.

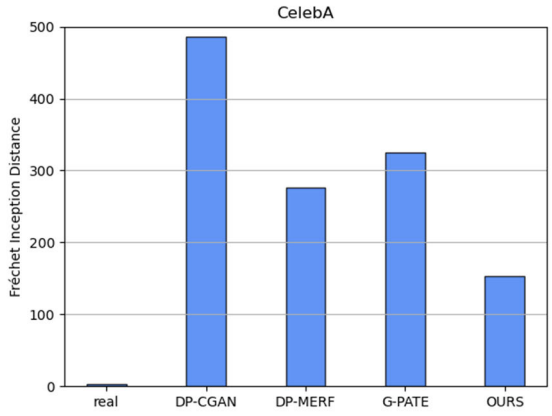

FIGURE 9. Fréchet Inception Distance of real and synthetic data on CelebA.

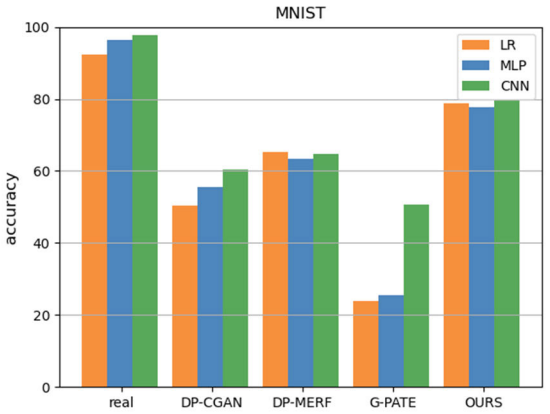

FIGURE 10. Classification accuracy for classification tasks on MNIST.

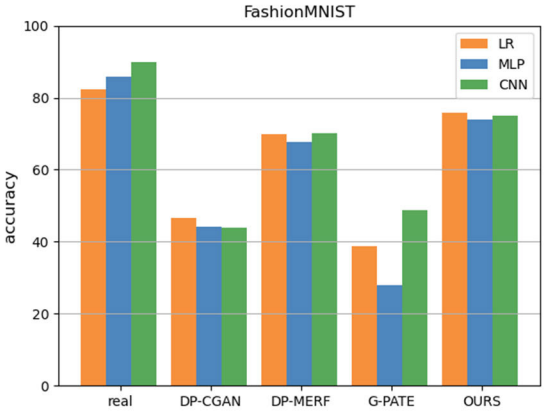

FIGURE 11. Classification accuracy for classification tasks on Fashion MNIST.

by calculating the Kullback-Leibler (KL) divergence between the conditional distribution of generated images and the class distribution. A higher Inception Score indicates better diversity and realism of generated images across different categories. And Fréchet Inception Distance (FID) [33], which is used to assess the similarity between the distribution of real images and generated images produced by a generative model. It measures the distance between two multivariate Gaussian distributions, one representing the feature representations of real images and the other representing the feature representations of generated images. A lower FID score indicates better quality and higher similarity between the generated images and real images. The FID is widely used in evaluating and comparing the performance of different generative models. Utility measured by classification accuracy. We train three classifiers on 30k privately generated images and evaluate the prediction accuracy on the test set. We consider Logistic Regression (LR), multi-layer perceptron (MLP), and convolutional neural networks (CNN) classifiers.

C. HYPERPARAMETERS

In this paper, we apply the cosine noise scheme which is used to determine the weight of the noise added in the forward diffusion process. We set timesteps to 1000, the number of iterations during model training to 100, and the batch size to 64. According to work [6], this paper sets the privacy leakage probability δ to 10^{-5}.

TABLE 1. Impact of different schemas.

Schema	Parameter Group	FID
Linear Decay	dr=0.3, lr_initial=1e-3	134
Exponential Decay	dr=0.99, lr_initial =1e-3	141
Logarithmic Decay	dr=0.002 lr_initial =1e-3	127
No Decay	1e-2	492
	1e-3	187
	1e-4	154
	1e-5	174
	1e-6	328

D. BASELINES

We consider the following state-of-the-art methods: DP-CGAN [7], DP-MERF [12] and G-PATE [11]. For a fair comparison, we evaluate all methods with a privacy parameter of $(\varepsilon, \delta) = (10, 10^{-5})$ via 30k generated images.

Fig. 3 shows the visual effect of image generation by DPDM on MNIST and Fashion-MNIST when the privacy budget ε is 10.

Fig. 4, Fig. 5 and Fig. 6 show the Inception Score of the synthetic images using four generative models and the real data from MNIST, Fashion-MNIST and CelebA. Fig. 7 and Fig. 8 and Fig. 9 show the Fréchet Inception Distance. We can find that the Inception Scores obtained by our algorithm are closer to the real dataset, and the difference between the value of the Inception Scores and that of the synthetic dataset generated by DDPM without noise is obviously smaller than the other algorithms. Among all SoTA baselines, the FID reflects that our algorithm generates better samples.

This paper refers to the semi-supervised classification algorithm, and uses the synthetic dataset to complete the image classification task. Specifically, analysts have a small amount of public labeled datasets and a large number of synthetic unlabeled data. The goal is to train a semi-supervised classifier with better performance by using labeled and unlabeled data. And then we get its classification accuracy on the MNIST and Fashion-MNIST test set. The experimental results are shown in Fig. 10 and Fig. 11. The classification accuracy of the DPDM framework proposed in this paper is close to the classification accuracy under the noiseless model, and it is better than the SoTA baselines.

As the learning rate impacts the model convergence, we also include how FID and generated images vary with different decay strategies. TABLE 1 reports the effects of different learning rate schemes on MNIST under $\left(10, 10^{-5}\right)$ −DP. And the results show that among a series of fixed learning rates, 1e-4 corresponds to the best image quality. Compared with the static learning rate, the dynamic learning rate has better performance, and the logic decay strategy is better than linear decay and exponential decay on FID.

VI. CONCLUSION

By combining differential privacy and diffusion models, this paper proposes a diffusion model (DPDM) for publishing image data with privacy protection. The trained diffusion model can be released directly for analysis tasks without worrying about disclosing the privacy of training data. Through experimental verifications, the results show that the algorithm in this paper can generate higher-quality image data. Compared with other works, the algorithm proposed in this paper also performs better in a series of metrics. For future work, we will consider trying different clipping methods to further reduce the privacy budget under the premise of retaining the high availability of images.

COMPETING INTEREST STATEMENT

The authors declare that they have no competing interests or other interests that might be perceived to influence the results reported in this paper.

REFERENCES

[1] H. Yin, A. Mallya, A. Vahdat, J. M. Alvarez, J. Kautz, and P. Molchanov, "See through gradients: Image batch recovery via GradInversion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 16332–16341, doi: 10.1109/CVPR46437.2021.01607.

[2] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Advances in Cryptology—EUROCRYPT 2006* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2006, pp. 486–503, doi: 10.1007/11761679_29.

[3] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends® Theor. Comput. Sci.*, vol. 9, pp. 211–407, Jan. 2013, doi: 10.1561/0400000042.

[4] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Advances in Cryptology—EUROCRYPT 2006* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2006, pp. 486–503, doi: 10.1007/11761679_29.

[5] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Vienna, Austria, Oct. 2016, pp. 308–318, doi: 10.1145/2976749.2978318.

[6] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," 2018, *arXiv:1802.06739*.

[7] R. Torkzadehmahani, P. Kairouz, and B. Paten, "DP-CGAN: Differentially private synthetic data and label generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, doi: 10.1109/CVPRW.2019.00018.

[8] X. Zhang, S. Ji, and T. Wang, "Differentially private releasing via deep generative model (technical report)," Jan. 2018, *arXiv:1801.01594*.

[9] C. Xu, J. Ren, D. Zhang, Y. Zhang, Z. Qin, and K. Ren, "GANobfuscator: Mitigating information leakage under GAN via differential privacy," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 9, pp. 2358–2371, Sep. 2019, doi: 10.1109/TIFS.2019.2897874.

[10] J. Jordon, J. Yoon, and M. V. D. Schaar, "PATE-GAN: Generating synthetic data with differential privacy guarantees," presented at the Int. Conf. Learn. Represent., 2018. [Online]. Available: https://openreview.net/forum?id=S1zk9iRqF7

[11] Y. Long, S. Lin, Z. Yang, C. A. Gunter, H. Liu, and B. Li, "Scalable differentially private data generation via private aggregation of teacher ensembles," presented at the Int. Conf. Learn. Represent., Sep. 2019. [Online]. Available: https://openreview.net/forum?id= Hkl6i0EFPH

[12] F. Harder, K. Adamczewski, and M. Park, "DP-MERF: Differentially private mean embeddings with random features for practical privacy-preserving data generation," 2020, *arXiv:2002.11603*.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *J. Jpn. Soc. Fuzzy Theory Intell. Inform.*, vol. 10, p. 177, Oct. 2017, doi: 10.3156/JSOFT.29.5_177_2.

[14] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[15] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*.
[16] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," 2015, *arXiv:1503.03585*.
[17] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," 2021, *arXiv:2102.09672*.
[18] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.
[19] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy*, Oakland, CA, USA, May 2008, doi: 10.1109/SP.2008.33.
[20] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, Oct. 2002, doi: 10.1142/S0218488502001648.
[21] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, Atlanta, GA, USA, 2006, pp. 1–6, doi: 10.1109/ICDE.2006.1.
[22] 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, doi: 10.1109/ICDE.2007.367856.
[23] S. Goldwasser and S. Micali, "Probabilistic encryption," *J. Comput. Syst. Sci.*, vol. 28, pp. 270–299, Apr. 1984, doi: 10.1016/0022-0000(84)90070-9.
[24] M. Diaz, H. Wang, F. P. Calmon, and L. Sankar, "On the robustness of information-theoretic privacy measures and mechanisms," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 1949–1978, Apr. 2020.
[25] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
[26] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Neural Inf. Process. Syst.*, Jan. 2020, pp. 1–13.
[27] I. Mironov, "Rényi differential privacy," in *Proc. IEEE 30th Comput. Secur. Found. Symp. (CSF)*, Santa Barbara, CA, USA, Aug. 2017, pp. 263–275, doi: 10.1109/CSF.2017.11.
[28] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.
[29] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.
[30] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3730–3738, doi: 10.1109/ICCV.2015.425.
[31] C. Li et al., "ALICE: Towards understanding adversarial learning for joint distribution matching," in *Proc. Neural Inf. Process. Syst.*, Sep. 2017, pp. 5495–5503.
[32] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," 2016, *arXiv:1606.03498*.
[33] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Neural Inf. Process. Syst.*, Jan. 2017, pp. 6629–6640.

ZHIGUANG CHU is currently pursuing the Ph.D. degree with the Beijing University of Technology. He has published articles in several journals. His current research interests include data security and privacy protection. After studying in this field for many years, he likes to participate in innovative competitions and won awards.

JINGSHA HE received the bachelor's degree in computer science from Xi'an Jiaotong University, China, and the master's and Ph.D. degrees in computer engineering from the University of Maryland, College Park, MD, USA. He worked for several multinational companies in the USA, including IBM Corporation, MCI Communications Corporation, and Fujitsu Laboratories. He is currently a Professor with the Faculty of Information Technology, Beijing University of Technology (BJUT), Beijing. He has published more than ten articles. He holds 12 U.S. patents. Since August 2003, he has published more than 300 articles in scholarly journals and international conferences. He also holds more than 84 patents and 57 software copyrights in China and authored nine books. He was a principal investigator of more than 40 research and development projects. His current research interests include information security, wireless networks, and digital.

DONGDONG PENG received the bachelor's degree from the Liaoning University of Technology, Jinzhou, Liaoning, China, in 2021, where he is currently pursuing the master's degree. His current research interests include information security, deep learning, and the Internet of Things data security. He is a member of Key Laboratory of Security for Network and Data in Industrial Internet of Liaoning Province, China.

XING ZHANG is currently a Professor with the School of Electronic and Information Engineering, Liaoning University of Technology. He has authored more than 30 peer-reviewed articles in leading journals and conference proceedings. His current research interests include network architecture and protocol, information security, and the Internet of Things technology. He serves as a member for ACM and CCF. He is the Director of Key Laboratory of Security for Network and Data in Industrial Internet of Liaoning Province, China.

NAFEI ZHU received the B.S. and M.S. degrees from Central South University, China, in 2003 and 2006, respectively, and the Ph.D. degree in computer science and technology from the Beijing University of Technology, Beijing, China, in 2012. From 2015 to 2017, she was a Postdoctoral Research Fellow with the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing. She is currently an Associate Professor with the Faculty of Information Technology, Beijing University of Technology. She has published more than 20 research articles in scholarly journals and international conferences. Her current research interests include information security and privacy, wireless communications, and network measurement.