

## Privacy Protection in Generative Model

Haihan Gao<sup>1</sup>, Xiaolu Chen<sup>2</sup>, Li Zhang<sup>3</sup>

University of Science and Technology of China

2022.10.26

<sup>&</sup>lt;sup>1</sup>SA22011017

<sup>&</sup>lt;sup>2</sup>SA22221005 <sup>3</sup>SA22221064

# Outline



# Data for machine learning

## ■ Age of big data

machine learning becoming a hit require a large amount of data from real world

### ■ Privacy data

like human face and fingerprint... sensitive and individual but of high practical value a tradeoff between data diversity and data utilization



# Generative model for privacy protection

### Generative model

adding noise to real data without exposing the original figure little significance if only copy input to output

#### Related work

## DP-SGD[?]

effective in generating high-dimensional sanitized data difficult implement

#### PATE

only train generators with DP guarantees high privacy costs

### Fed-Avg GAN

provide user-level DP guarantees merely works on decentralized data



## Introduction of GAN

#### Generator

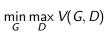
create fake data from random sample sample  $Z = (z_1, z_2, ...z_n)$  $\tilde{Z} = G(Z)$ 

#### Discriminator

discriminate between fake and real data  $D(X, G(Z)) = \begin{cases} 0, & \text{if input } \in G(Z) \\ 1, & \text{if input } \in X \end{cases}$ 

### ■ Task of two part

**Generator** G(Z) closer to X, the better **Discriminator** distinguish G(Z) and X more, the better





# Problem of generative model

- Most privacy-preserving training algorithms for neural network models is manipulating the gradient information generated during backpropagation.
- Methods
  Clipping the gradients (to bound sensitivity)
  Adding calibrated random noise (to introduce stochasticity)
- **Problem**:limited in shallow networks and fail to sufficiently capture the sample quality of the original data.



## Review DP

### ■ Neighboring dataset

D, D': differed with only a piece of data

## Differential Privacy

$$Pr[M(D) \in O] \le e^{\epsilon} Pr[M(D') \in O] + \delta$$

M:a random algorithm

O: output set

 $\epsilon$ : privacy budget smaller $\epsilon$ , M(D) and M(D') closer

 $\delta$ :disturbance



## Review DP

Suppose  $\mathcal{M}: \mathbb{R}^n \to X$ , X is called mapping Space and  $\mathcal{M}$  is called map from origin space into mapping space. For two dataset x and x', they only has one-item difference.We call  $\mathcal{M}$  satisfies  $(\epsilon, \delta)$ -privacy if following statement is true

$$Pr[\mathcal{M}(x) \in S] \le e^{\epsilon} Pr(\mathcal{M}(x') \in S) + \delta$$
 (1)

What is  $\mathcal{M}$ ,input space and mapping space in this scenario?



## Review DP

According the define of [?],  $\mathcal{M}$  is the generative model.Its input space is consisted with training data and its mapping space is consisted with its output. Take Conditional-VAE for an example.Suppose training data takes with following formular

$$(X, y) = (x_i, y_i)_{i=1}^n$$
 (2)

it is consisted with n data samples and  $x_i \in \mathbb{R}^n$ ,  $y_i \in \mathbb{R}$ ,  $x_i$  is the data(such as an image) and  $y_i$  is the label of the image.  $\mathcal{M}$  is a model which generate new data from existing data and label.



# Privacy define

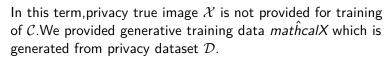
Possibility is hard to quantify for NN model.So we take several ways to varify our model satisify  $(\epsilon, \delta)$  privacy



## From attackers

One of the goal of the attacker is to varify whether a data is from or not from the dataset. If the privacy dataset  $\mathcal D$  is revealed. We can build the attacker from  $\mathcal D$ 

- sample k data  $x \in \mathcal{D}$  and sample k noise from  $\mathbb{R}^n$  but  $\notin \mathcal{D}$ . Denoted as  $X = x_1, x_2, \cdots, x_k$  and  $Y = y_1, y_2, \cdots, y_k$
- X is true data,we attack label True to all its data and Y is fake data which randomly sample from  $\mathbb{R}^n$ ,so we attach False label into Y,mixture X, Y and its labels T together
- build a classifier C, which takes true/fake image as input and output the Possibility of whether the input pattern is belonged to True/Generated Data. Train the model with supervised learning method





## For attackers

suppose P is the Possibility attacker could correctly varify the real data through generative data. We define privacy loss

$$privacy loss = \frac{P - 0.5}{0.5} \tag{3}$$



# Utility experiment

use generative data for image classifier task. We train the downstream image classifier model with Training data  $T_{train}$  and validate the training accuracy with testing data  $T_{train}$ . There are four condition for utility measurement

Test. There are rour condition for dentry measurement.					
accuracy test data train data	privacy data	generative data			
privacy data	$p_1$	$p_2$			
generative data	<i>p</i> <sub>3</sub>	$p_4$			

We call  $p_1 - p_2$  accuracy loss from the generative process



## Work schedule

- Haihan Gao: Algorithm design and implementation of main model and add privacy into whole model
- Xiaolu Chen: Research of previous work and implementation of Attackers
- Li Zhang: Report writing, research of privacy attack and implementation of utility network



