AutoGAN-based Dimension Reduction for Privacy Preservation

Hung Nguyen[a], Di Zhuang[a], Peiyuan Wu[b], Morris Chang[a]

[a] *University of South Florida, USA*
[b] *National Taiwan University, Taiwan*

Abstract

Exploiting data and concurrently protecting sensitive information to whom data belongs is an emerging research area in data mining. Several methods have been introduced to protect individual privacy and at the same time maximize data utility. Unfortunately, existing techniques such as differential privacy are not effectively protecting data owner privacy in the scenarios using visualizable data (e.g., images, videos). Furthermore, such techniques usually result in low performance with a high number of queries. To address these problems, we propose a dimension reduction-based method for privacy preservation. This method generates dimensionally-reduced data for performing machine learning tasks and prevents a strong adversary from reconstructing the original data. In this paper, we first introduce a theoretical tool to evaluate dimension reduction-based privacy preserving mechanisms, then propose a non-linear dimension reduction framework using state-of-the-art neural network structures for privacy preservation. In the experiments, we test our method on popular face image datasets and show that our method can retain data utility and resist data reconstruction, thus protecting privacy.

Keywords: Generative Adversarial Nets, Auto-encoder, neural-network, privacy preservation, dimension reduction, access control.

1. Introduction

The applications of machine learning (ML) such as an on-line system collecting data from multiple data owners are raising privacy issues. Many types of user data are being collected in a smart city such as patient records, salary information, biological characteristics, Internet access history, personal images. These types of data can be widely used in daily recommendation systems, business data analysis, or a disease prediction system. However, collecting and using such data might raise privacy issues for individuals who contribute their sensitive information. Considering the multi-level access control system of a company using biometric recognition (e.g., facial images) for granting permission to access data resources, its staff members may concern their facial information being vulnerable to adversaries. The utility of face features can be effectively used in machine learning tasks for authentication purposes. However, leaking the face images might lead to a privacy problem such as the utilization of face images to determine the members' identity by an adversary. In this study, we consider an access control system collecting dimensionally-reduced face images of staff members to perform authentication task and to provide permission for members who would like to access the data resources. We propose a non-linear dimension reduction framework to decrease data dimension for the authentication purpose mentioned above and prevent an adversary from reconstructing their face images.

Several tools and methods have been developed to preserve the privacy in machine learning, such as homomorphic encryption [1–3], secure multi-party computing [4, 5], differential privacy (DP) [6–10], compressive privacy [11–15]. Differential privacy-based methods aim at preventing leaking individual information caused by queries. However, they are not designed for a large number of queries since they require adding a huge amount of noise for privacy preserving, thus significantly decreasing the ability to learn meaningful information from data. Homomorphic encryption-based methods can be used to privately evaluate a function over the encrypted data by a third party without accessing to the plain-text data. Consequently, the privacy of data owners can be protected. However, due to the high computational cost and time consumption, they may not work with a very large dataset normally required in ML applications. To tackle these problems, we introduce a theoretical tool for dimension reduction privacy evaluation $\epsilon - DR$ *privacy*. This tool evaluates the distance between original data and reconstructed data of a dimension reduction (DR) mechanism. Because in our tool high distance yields high level of privacy, DR mechanisms are encouraged to enlarge the distance. While DP relies on inference uncertainty to protect sensitive data, $\epsilon - DR$ privacy is built on reconstruction error to evaluate privacy. Unlike DP,

Email addresses: nsh@mail.usf.edu (Hung Nguyen), zhuangdi1990@gmail.com (Di Zhuang), peiyuanwu@ntu.edu.tw (Peiyuan Wu), morrisjchang@gmail.com (Morris Chang)

$\epsilon - DR$ privacy is not negatively impacted by the number of queries. Furthermore, in section 6 we recommend a privacy-preserving system Autoencoder Generative Adversarial Nets-based Dimension Reduction Privacy (AutoGAN-DRP) for enhancing data owner privacy and preserving data utility. The *utility* herein is evaluated via machine learning task (e.g., classification) performance. Generative Adversarial Privacy (GAP) [12] is a similar method utilizing the minimax algorithm of Generative Adversarial Nets to preserve privacy and to keep utility on image datasets. GAP perturbs data within a specific l_2 distance constraint between original and perturbed data to distort private class labels and at the same time preserve non-private class labels. However, it does not protect an image itself, and an adversary can intuitively infer private label (e.g., identity) from an image. Further, high correlation between private labels and non-private labels might result in a dramatic decrease in accuracy when privacy level is increased. In contrast, our method protects an image by compressing it into a few dimensions vector and then transferring without exposing the original image. Our framework can be applied to different types of data and used in several practical applications without heavy computation of encryption and impact of query number. The proposed framework can be applied directly to the access control system mentioned above. More elaboratively, face images are locally collected, nonlinearly compressed to achieve DR, and sent to the authentication center. The authentication server then performs authentication tasks on the dimensionally-reduced data. We assume the authentication server is semi-honest, that is to say it does not deviate from authenticating protocols while being curious about a specific user's identity. The DR framework prevents a strong adversary who obtains a training dataset and a transformation model from reconstructing a user's face image after he or she sends an authentication request. We perform several experiments on benchmark datasets to show our method performance. The experiments illustrate that our method can achieve high accuracy with very small number of reduced dimensions and satisfy $\epsilon - DR$ privacy.

Our work has two main contributions:

- To analytically support privacy guarantee, we introduce a theoretical tool $\epsilon - DR$ *privacy* for privacy preserving mechanism evaluation.
- We propose a non-linear dimension reduction framework for privacy preservation motivated by Generative Adversarial Nets [16] and Auto-encoder Nets [17].

The rest of our paper is organized as follows. Section II summarizes state-of-the-art privacy preservation machine learning (PPML) techniques. Section III reviews knowledge of deep learning methods including generative adversarial neural nets and Auto-encoder. Section IV describes the privacy problem through a scenario of leaking individual information in a facial recognition access control system. In section V, we introduce the definition of $\epsilon - DR$ privacy, a tool to evaluate DR-based privacy preserving mechanisms. Section VI presents our framework AutoGAN-based Dimension Reduction for Privacy Preservation. Section VII displays the experiment results. Finally, the conclusion and future work are mentioned in section VIII.

2. Literature Review

Cryptographic approach: This approach usually applies to the scenarios where the data owners do not wish to expose their plain-text sensitive data while asking for machine learning services from a third-party. The most common tool used in this approach is fully homomorphic encryption that supports multiplication and addition operations over encrypted data, which enable the ability to perform a more complex function. However, the high cost of the multiplicative homomorphic operations renders it difficult to be applicable on machine learning tasks. In order to avoid multiplicative homomorphic operations, additive homomorphic encryption schemes are more widely used in privacy preserving machine learning (PPML). However, the limitation of the computational capacity in additive homomorphic schemes narrows the ability to apply on particular ML techniques. Thus, such additive homomorphic encryption-based methods in [1, 2] are only applicable to simple machine learning algorithms such as decision tree and naive bayes. In Hesamifard's work [3],the fully homomorphic encryption is applied to perform deep neural networks over encrypted data, where the non-linear activation functions are approximated by polynomials. In secure multi-party computing (SMC), multiple parties collaborate to compute functions without revealing plain-text to other parties. A widely-used tool in SMC is garbled circuit [4], a cryptographic protocol carefully designed for two-party computation, in which they can jointly evaluate a function over their sensitive data without the trust of each other. In [18], Mohammad introduced a SMC protocol for principle component analysis (PCA) which is a hybrid system utilizing additive homomorphic and garbled circuit. In secret sharing techniques [5], a secret **s** is distributed over multiple pieces **n** also called *shares*, where the secret can only be recovered by a sufficient amount of **t** *shares*. A good review of secret sharing-based techniques and encryption-based techniques for PPML is given in [19]. Although these encryption-based techniques can protect the privacy in particular scenarios, computational cost is a significant concern. Furthermore, as [19] elaborated, the high communication cost also poses a big concern for both techniques.

Non-Cryptographic approach: Differential Privacy (DP) [20] adds noise to prevent membership inference attack by providing a tool as follows. A mechanism M satisfies ϵ-differential privacy if for any two neighbor databases D and D' which differ by at most one element and any subset S of the output space of M satisfies $Pr[M(D) \in S] \leq e^{\epsilon}.Pr[M(D') \in S]$. The similarity of query outputs (whether or not a certain record is included in the dataset) protects a member information from membership inference attack. The *similarity* is guaranteed by the parameter ϵ in a mechanism in which the smaller ϵ provides a better level of privacy preservation.[6], [7] propose methods to guarantee ϵ-differential privacy by adding noise to outcome of the weights $w^* = w + \eta$, where η drawn from Laplacian distribution and adding noise to the objective function of logistic regression or linear regression models. [8], [9] satisfy differential privacy by adding noise to the objective function while training a deep neural network using stochastic gradient descent as the optimization algorithm. There are also existing works proposing differential privacy dimension reduction. Dimension reduction is an important process before data goes through machine learning algorithm. Hence, one can guarantee ϵ-differential privacy by perturbing dimension reduction outcome. Principal component analysis (PCA) whose output is a set of eigenvectors is a popular method in dimension reduction. The original data is then represented by its projection on those eigenvectors, which keeps the largest variance of the data. One can reduce the data dimension by eliminating insignificant eigenvectors which contain less variance, and apply noise on the outcome to achieve differential privacy[10]. However, the downside of these methods is that they are designed for specific mechanisms and datasets and not working well with the others. For example, record-level differential privacy is not effectively used with image dataset as shown in [21]. Also, the amount of added noise is accumulative based on the number of queries so that this approach usually leads to low accuracy results with a high number of queries.

3. Preliminaries

To enhance the distance between original and reconstructed data in our DR system, we utilize the structure of Generative Adversarial Network (GAN) [16] for data manipulation and deep Auto-encoder [17] as a reconstruction method. The following sections briefly review Auto-encoder and GAN.

3.1. Auto-encoder

Auto-encoder, an artificial neural network, is aimed at learning a lower dimension representation of an unsupervised data or generative models. Auto-encoder can be used for denoising image data and reducing dimension. Auto-encoder can be implemented by two fully connected neural network components: encoder and decoder. The encoder and decoder perform reverse operations. The encoder input is normally original data while the decoder output is expected to be similar to the input data. The middle layer which has smaller number of neurons than that of the input is extracted as a representation of dimensionally-reduced data. The auto-encoder training process can be described as a minimization problem of a loss function:

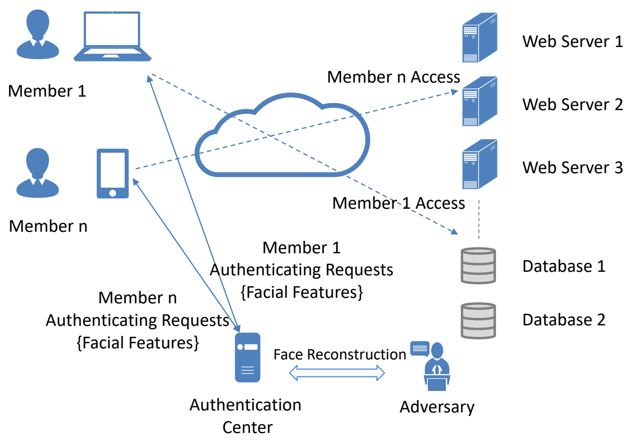


Figure 1: Attack Model

$$\mathcal{L}(x, g(f(x)))$$

where x is the data and g(.) is a function of the framework which could be the mean square error between input and output.

3.2. GAN

Generative Adversarial Nets is aimed at approximating distribution p_d of a dataset via generative models over data x. GAN simultaneously trains two generative models G and D in which G inputs are sampled from a prior distribution $p_z(z)$. G generates fake samples similar to the real samples. At the same time, D is trained to differentiate between fake samples and real samples and send feedback to G for improvement. GAN can be formed as a two-player minimax game with value function V(G,D):

$$\underset{G,D}{min\ max}\ V(G,D) = E_{x\ p_d(x)}[log(D(x))] + E_{z\ p_z(z)}[log(1 - D(G(z)))]$$

In practice, the two GAN components, *generator* and *discriminator* are built on two fully connected neural networks. The loss function of G is to reduce the D accuracy. Meanwhile, the loss function of D is to increase the accuracy of differentiating fake samples from real samples. These two components are trained until the discriminator cannot distinguish between generated samples and real samples.

4. Problem

4.1. Problem statement

We introduce the problem through the practical scenario mentioned in Section 1. Staff members in a com-

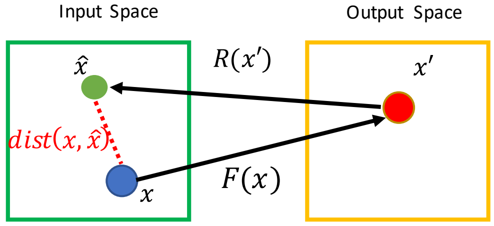


Figure 2: DR projection and reconstruction.

pany request access to its data resources, such as websites and data servers with different privileges through a face recognition access control system. The system collects each member's facial features and sends them to a central server to determine his/her access eligibility. We consider that the system has three levels of privilege (i.e., single-level, four-level, eight-level) corresponding to three member groups. We assume the authentication server is semi-honest (it obeys the work procedure but might be used to infer personal information). An adversary in the authentication center can reconstruct the face features to achieve plain-text face images and determine members' identity.

4.2. Threat model

In the above scenario, we consider that a very strong adversary who has access to the model and training dataset attempts to reconstruct the original face images for inferring a specific member's identity. Our attack model can be represented in Figure 1. The adversary utilizes training data and facial features to identify a member identity by reconstructing the original face using Auto-encoder nets. Rather than using fully connected neural network, we implement Auto-encoder on convolutional neural network to create an effective model for image dataset. Our goal is to design a data dimension reduction method for reducing data dimension and resisting full reconstruction of original data.

5. ϵ-Dimension Reduction Privacy (ϵ-DR Privacy)

In this section, we introduce the Dimension Reduction Privacy (DR-Privacy), and define a formal definition of the ϵ-DR Privacy to mathematically quantify/evaluate the mechanisms designed to preserve the DR-Privacy via dimension reduction. The DR-Privacy aims to achieve privacy-preserving via dimension reduction, which refers to transforming the data into a lower dimensional subspace, such that concealing the private information while preserving the underlying probabilistic characteristics, which can be utilized for machine learning purposes. To quantify the DR-Privacy and guide us to design such DR functions, we define ϵ-DR Privacy as follows.

Definition 1: (ϵ-DR Privacy) A Dimension Reduction Function $F(\cdot)$ satisfies $\epsilon - DR$ Privacy if for each i.i.d. m-dimension input sample x drawn from the same distribution D, and for a certain distance measure $dist(\cdot)$, we have

$$E[|dist(x,\hat{x})|] \geq \epsilon \tag{1}$$

where $\epsilon \geq 0$, $x' = F(x)$, $\hat{x} = R(x')$, and $R(\cdot)$ is the Reconstruction Function.

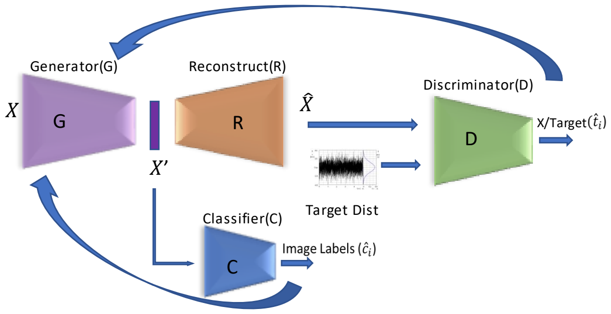


Figure 3: AutoGAN-DRP

For instance, as shown in Fig. 2, given original data x, our framework utilizes certain dimension reduction function $F(x)$ to transform the original data x into the transformed data x'. The adversaries aim to design a corresponding reconstruction function $R(x')$ such that the reconstructed data $\hat{x}$ would be closed/similar to the original data x. DR-Privacy aims to design/develop such dimension reduction functions, that the distance between the original data and its reconstructed data would be large enough to protect the privacy of the data owner.

6. Methodology

6.1. AutoGAN-based Dimension Reduction for Privacy Preserving (AutoGAN-DRP)

To tackle the problem, we propose a deep learning framework for transforming face images to lower dimension vectors which are hard to be clearly reconstructed. The dimensionally-reduced data can be sent to the authentication server as an authentication request. We consider an adversarial as a re-constructor implemented by a Deep Auto-encoder structure. To prevent fully reconstructing images, the framework utilizes the discriminator in GAN [16] to direct reconstructed data to a desired target distribution. In this work, the target distribution is sampled from Gaussian distribution and the mean is the average of the whole training data. After the transformation projects the data into a lower dimension domain, the re-constructor can only partially reconstruct the data. Therefore, the adversary might not be able to recognize an individual's identity. To maintain data utility, we use feedback from a classifier. The entire framework can enlarge the distance between original data and its reconstruction to preserve individual privacy and retain significant data information. The dimensionally-reduced transformation model is extracted from the framework and provided to clients for reducing their facial image dimensions. The classification model will be used in an authentication center that classifies whether an authentication request is valid to have access {1} or not {0}.

Algorithm 1 Algorithm for stochastic gradient descent training of ϵ -DR privacy.

```
Input: Training dataset $D$.
    Parameter: learning rate $\alpha_r, \alpha_d, \alpha_c, \alpha_g$, training steps $n_r, n_d, n_c, n_g$
    A constraint for $\epsilon - DR$
Output: Transformation Model
    Initialization.
for number of global training iteration do
    Randomly, sample a mini batch from target distribution and label $\boldsymbol{t}$.
    Randomly, sample mini batch of data $\boldsymbol{x}$ and corresponding label $\boldsymbol{y}$
    for $i = 0$ to $n_r$ iterations do
        Update the Reconstruction and Generator:
        $\varphi_{i+1} = \varphi_i - \alpha_r \nabla_\varphi \mathcal{L}_R(\varphi_i, \boldsymbol{x})$
        $\theta_{l+1} = \theta_l - \alpha_r \nabla_\varphi \mathcal{L}_R(\varphi_i, \boldsymbol{x})$
    end for
    for $j = 0$ to $n_d$ iterations do
        Update the Discriminator parameter:
        $\omega_{j+1} = \omega_j - \alpha_d \nabla_\omega \mathcal{L}_D(\omega_j, \boldsymbol{x}, \boldsymbol{t})$
    end for
    for $k = 0$ to $n_c$ iterations do
        Update the Classifier parameter:
        $\phi_{k+1} = \phi_k - \alpha_c \nabla_\phi \mathcal{L}_C(\phi_k, \boldsymbol{x}, \boldsymbol{y})$
    end for
    for $l = 0$ to $n_g$ iterations do
        Update the Generator parameter:
        $\theta_{l+1} = \theta_l - \alpha_g \nabla_\theta \mathcal{L}_G(\theta_l, \boldsymbol{x}, \boldsymbol{t}, \boldsymbol{y})$
    end for
end for
return
```

We formulate the problem as follows: Let D be the public training dataset. (x_i, y_i) are the i samples in the dataset which each sample x_i has d features and a ground truth label y_i. The system is aimed at learning a dimension reduction transformation $F(.)$ which transforms the data from d dimensions to d' dimensions in which $d' << d$. Let D' be the dataset in lower dimension domain. The dimensionally-reduced data should keep significant information to work with different types of machine learning tasks and should resist against the reconstruction or inference from data owner information.

Our proposed framework can learn a DR function $F(.)$ that preserves privacy at certain value of ϵ evaluated by $\epsilon - DR$ privacy. The larger distance implies higher level of privacy. Figure-3 represents our learning system in which D' is generated from a Generator G. D' can be classified by a classifier C and can resist data reconstruction of an aggressive attack implemented by a trainable Reconstructor R. We use a binary classifier for single-level authentication system and multi-class classifier for multi-level authentication system. The Discriminator D is a neural network working with G as a minimax game. The Discriminator aims to differentiate the reconstructed data from a target distribution and send feedback to Generator. The Generator is trained so that it directs reconstructed data distribution close to the target distribution to ensure a distance between reconstructed and original data. To enlarge the distance, the selected target distribution should be different from the original data distribution. The problem becomes finding an optimal solution for the Generator, as shown in (2):

$$\begin{aligned}\theta^* = \underset{\theta}{arg\ min}\ (\alpha \underset{\phi}{arg\ min}\ \mathcal{L}_C - \beta \underset{\omega}{arg\ min}\ \mathcal{L}_D \\ - \gamma \underset{\varphi}{arg\ min}\ \mathcal{L}_R + \mathcal{C}(\epsilon))\end{aligned} \tag{2}$$

Where α, β, γ are weights of components in the objective function and freely tuned. $\mathcal{C}(\epsilon)$ is a constraint function with respect to hyper-parameter ϵ.

The re-constructor plays its role as an aggressive adversary attempting to reconstruct original data by training R using known data. The loss function of R is the mean square error of original training data and reconstructed data, as displayed in (3):

$$\mathcal{L}_R = \sum_i^n (x_i - \hat{x}_i)^2 \tag{3}$$

The classifier C keeps the performance of the classification task in a lower dimension domain and sends feedback to G. The classifier loss function (4) is defined by a cross entropy of the class target y and predicted class $\hat{y}$. $\mathcal{L}_D$ (5) is the cross-entropy loss of the Discriminator.

$$\mathcal{L}_C = -\sum_{i=1}^n \sum_{j=1}^m y_{ij} \log(\hat{y}_{ij}) \quad y, \hat{y} \in \{0, 1, .., m-1\} \tag{4}$$

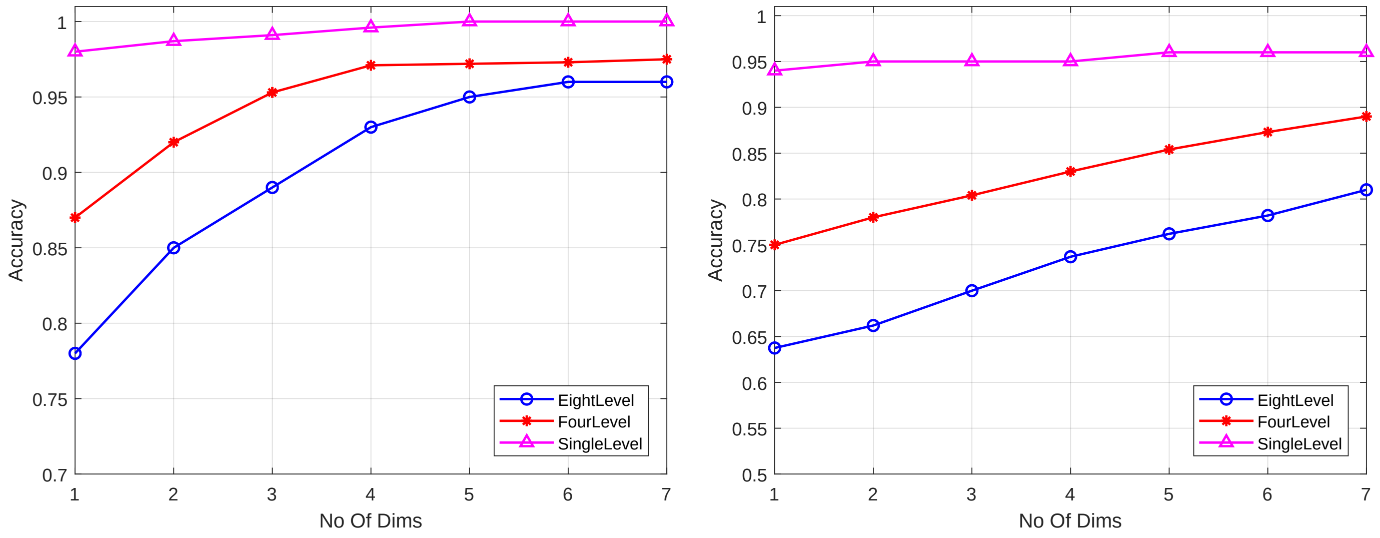
$$\mathcal{L}_D = -\sum_i^n (t_i \log(\hat{t}_i) + (1 - t_i) \log(1 - \hat{t}_i)) \quad t, \hat{t} \in \{0, 1\} \tag{5}$$

Where m denotes the number of classes and n denotes the number of samples.

6.2. Optimization With Constraint

In order to meet a certain level of distance, we use ϵ as a hyper-parameter to be tuned in the system. We use an constrained optimization method to put a constraint ϵ on the Generator as a part of its objective function. We choose the exterior method as our solution. Considering a constrained problem:

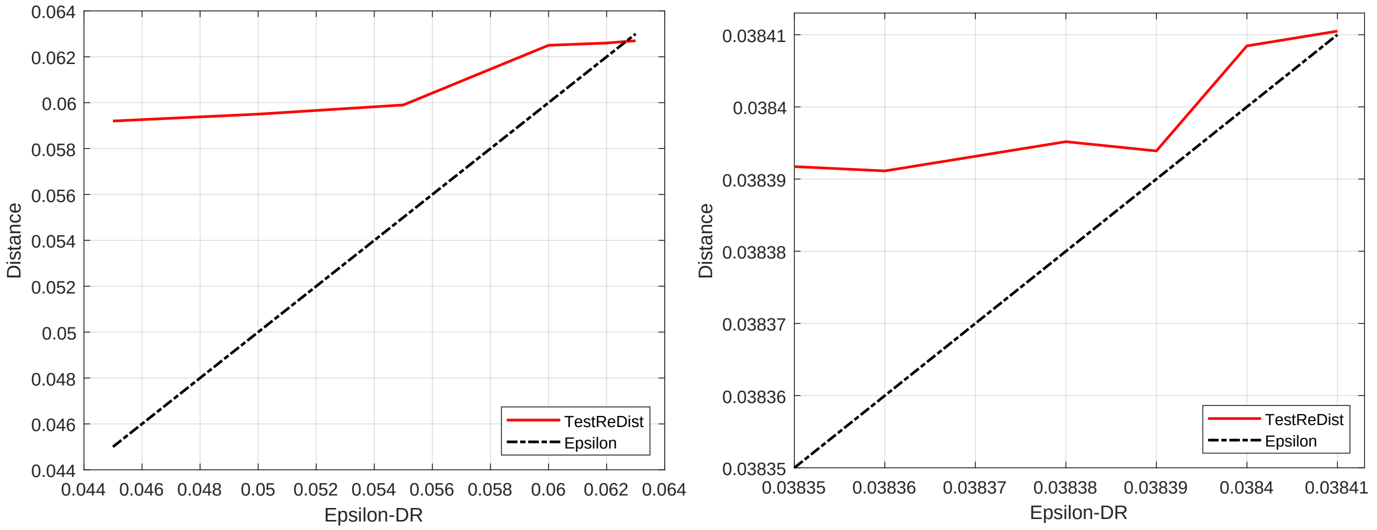
$$\underset{x}{minimize} f(x) \qquad s.t\, x \in \Omega$$



(a) Yale_B

(b) AT&T

Figure 4: Accuracy for Different Number of Reduced Dimensions



(a) Yale_B

(b) AT&T

Figure 5: Distance Measurement Result { 7 dimensions, Single-Level}

It could be approximated as a unconstrained problem:

$$minimize f(x) + \gamma \mathcal{C}(x)$$

Where $\mathcal{C} : \mathbb{R}^n \rightarrow \mathbb{R}$ is a penalty function and γ is penalty parameter. $\mathcal{C}$ is continuous, $\mathcal{C}(x) \geq 0$ for all $x \in \mathbb{R}^n$ and $\mathcal{C}(x) = 0$ iff $x \in \Omega$. The penalty function in G loss function becomes

$$\mathcal{C} = \gamma.max(0, dist - \epsilon) \tag{6}$$

where *dist* is defined as the distance between original data and reconstructed data.

6.3. Algorithms

Algorithms 1 describes how we train the entire framework. The framework contains four components and they are trained one by one. First, similar to training an auto-encoder, the Re-constructor is trained in n_r iterations and Generator parameters are concurrently updated while other components are fixed. Second, the Discriminator is trained while the others fixed. Third, the Classifier is trained in n_c iterations. Fourth, the Generator is trained in n_g iterations. We repeat this training process until it reaches a number of global training iterations or the weights are very close to their previous state.

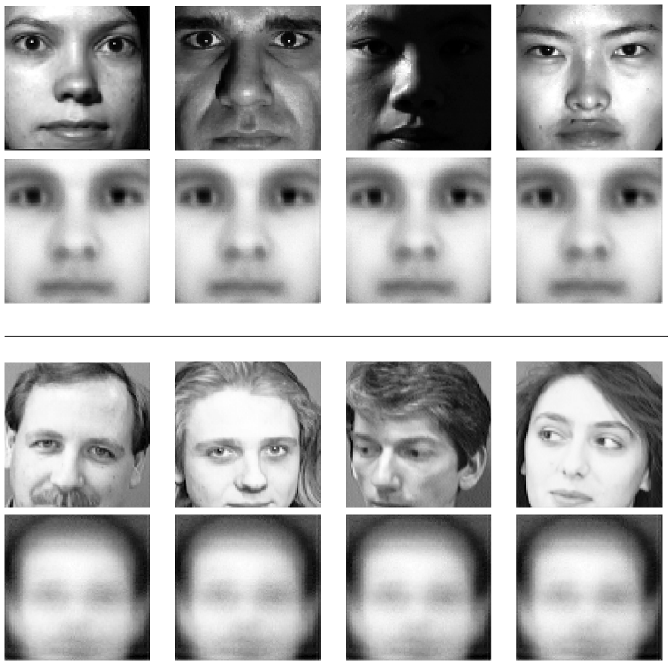


Figure 6: Visualization Samples of Original Images and Corresponding Reconstructions on Yale_B and AT&T
{SingleLevel, 7 dimensions }

7. Experiments and Discussion

In this section, we demonstrate our experiments on two popular supervised face image datasets: *the Extended Yale Face Database B* [22] and *AT&T* [23]. The Euclidean distance is used to measure the distance between original and reconstructed images: $dist(x, \hat{x}) = ||x - \hat{x}||^2$.

7.1. Experiment Setup

The Extended Yale Face Database B contains 2,470 grayscale images of 38 human subjects under different illumination conditions and their identity label. In this dataset, the image size is 168x192 pixels. In our experiment, we crop these images to 168x168 pixels. The AT&T dataset has 400 face images of 40 subjects. We crop the image size to 92x92 pixels. All pixel values are scaled to the range of [0,1]. We randomly select 20% of each subject's images for testing dataset.

The Generator and Re-constructor in Figure 3 are implemented as convolutional neural networks. Each of them has two convolutional layers and two fully connected hidden layers. Discriminator and Classifier are built on fully connected neural network with three hidden layers. Hyperbolic Tangent function is used as activation function for hidden layers. Each component is trained in 300 local iterations, and the entire system is trained in 1000 iterations with the learning rate of 0.01 for global loop. The target distribution is drawn from Gaussian distribution (with the covariance value of 0.05 and the mean is the average of the training data).

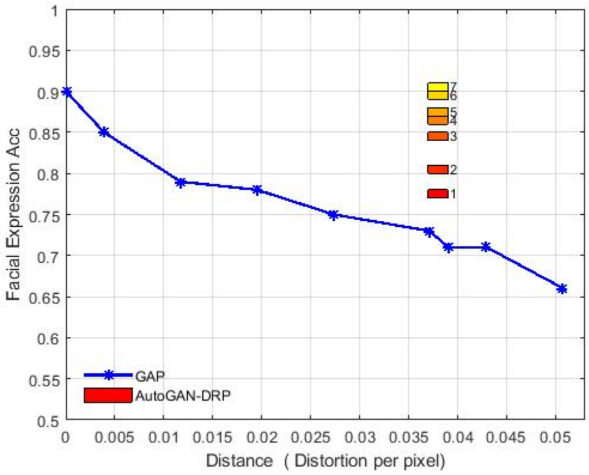


Figure 7: GENKI Facial Expression Accuracy Vs Distance using GAP and AutoGAN-DRP

For the single-level authentication system in the scenario, we consider half of the subjects in the dataset are valid to access the database system while the rest are invalid. We randomly divide the dataset into two groups of subject and labels their images to {1} or {0} depending on their access permission. For the cases of multi-level authentication system experiments, we divide the subjects into four groups and eight groups so that the authentication server becomes four-class and eight-class classifier respectively.

7.2. Utility

We use accuracy metric to evaluate the utility of dimensionally-reduced data. The testing dataset is tested with the classifier in our framework. Figure 4 illustrates the accuracies for different dimensions from one to seven. Overall, the accuracies improve when the dimension number increases. The accuracy result for single-level authentication system on YaleB starts from 97% with one dimension and reaches 100% with only five dimensions. As the dimension number is reduced from 28,224 (168x168) to 5, we can achieve a compression ratio of 5,644 yet achieve 100% of accuracy. In eight-level and four-level authentication, we can achieve accuracies of 97% and 96% with seven dimensions. Testing on AT&T dataset, we could achieve 77% of accuracy with only one dimension and 96% with seven dimensions for single-level authentication. This implies we could gain the compression ratio of 4,032 (from 92x92 to 7 dimensions) and maintain a high accuracy. These figures for four-level and eight-level authentication system at seven dimensions are 90% and 81% respectively. As shown in the figure 4, our method could achieve a very high result with a low number of dimension in terms of accuracy.

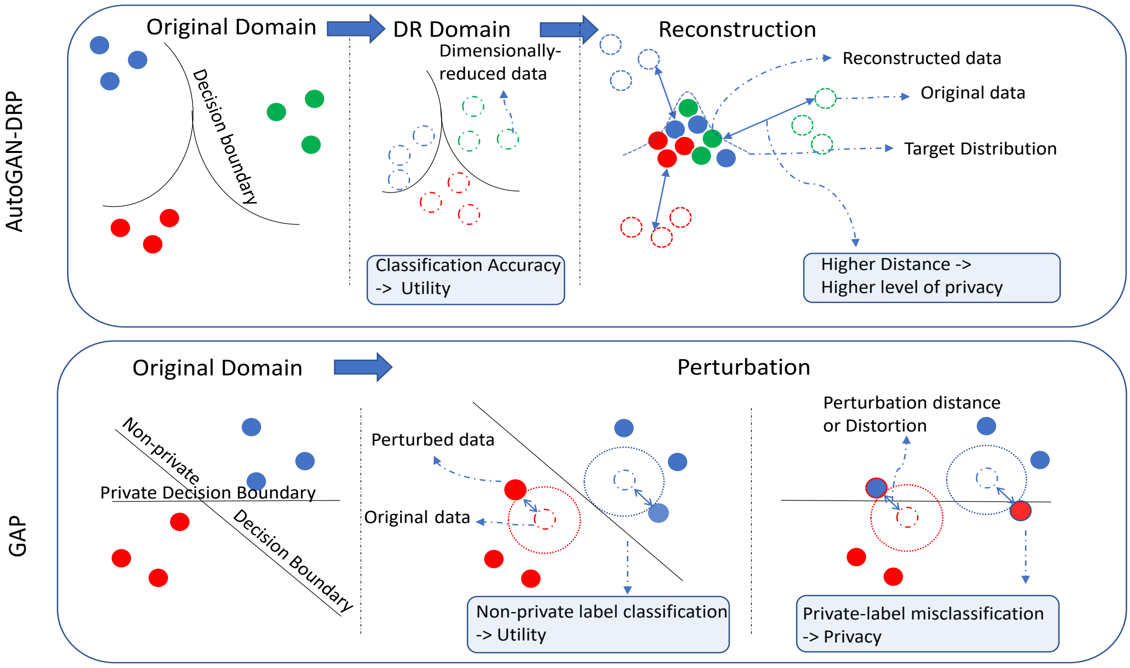

AutoGAN-DRP
Original Domain
Decision boundary
DR Domain
Dimensionally-reduced data
Classification Accuracy -> Utility
Reconstruction
Reconstructed data
Original data
Target Distribution
Higher Distance -> Higher level of privacy
GAP
Original Domain
Non-private Decision Boundary
Private Decision Boundary
Perturbation
Perturbed data
Original data
Non-private label classification -> Utility
Perturbation distance or Distortion
Private-label misclassification -> Privacy

Figure 8: AutoGAN-DRP Vs GAP

7.3. Privacy

Figure 5 illustrates the average distances between original images and reconstructed images (taken from the output of the Re-constructor) on testing data with different ϵ constraints for seven dimensions and single-level authentication. The achieved distances (red lines) are always larger than the hyper-parameter ϵ (black dotted lines) where ϵ is less than 0.063 for YaleB and 0.384 for AT&T. Due to the fact that the Re-constructor is trained using the training dataset (we consider the adversary can reach the model and the training data), our framework can only force the distance within a certain range. Specifically, the distances varies from 0.059 to 0.063 for YaleB and from 0.03839 to 0.03841 for AT&T. The intersection between the red line and the dotted black line points out the largest distance our framework can achieve. The method satisfies $\epsilon - DR$ privacy with ϵ values of 0.0384 for YaleB and 0.063 for AT&T. Figure 6 demonstrates some samples and their corresponding reconstructions in single-level authentication and seven dimensions. The reconstructed images are nearly identical, thus making it visually hard to recognize the identity of an individual.

8. Comparison to GAP[12]

In this section, we compare our framework with GAP. We attempt to visualize AutoGAN-DRP and GAP to highlight their similarities and differences. We also show our experiment results of the two methods on the same dataset. Our work shares many similarities with GAP, such as utilizing minimax algorithms of Generative Adversarial Nets, applying the state-of-the-art convolution neural nets for image datasets, considering l_2 norm distance (i.e., distortion in GAP, privacy measurement in AutoGAN-DRP) between the original data and possible exposed data. Specifically, both GAP and AutoGAN-DRP consider the difference between original and exposed images. This difference is understood as the *distortion* between original and perturbed images in GAP and the *distance* between original and reconstructed images in AutoGAN-DRP. Both of the *distance* and *distortion* are computed with $l2$ norm distance. In this context, the distance and distortion refer to the same measurement and have the same meaning. To be consistent, we use the term *distance* to present this measurement in the rest of this section.

Figure 8 illustrates the visualization of AutoGAN-DRP and GAP. In AutoGAN-DRP, the privacy is assessed based on how well an adversary can reconstruct the original data and measured by the distance between original and reconstructed data. The dimensionally-reduced data is reconstructed using the state-of-the-art neural network (an Auto-encoder). The larger the distance is, the more privacy can be achieved. Further, if the reconstructed images are blurry, privacy can be preserved since it is hard to visually determine an individual identity. The data utility is quantified by the accuracy of identity classification task over dimensionally-reduced data which captures the most significant data information. Meanwhile, GAP perturbs images with a certain distortion constraint to achieve privacy. It evaluates data utility by the classification accuracy of non-private label and assesses privacy by the classification accuracy of private label. Similar to AutoGAN-DRP, the high distortion is most likely to yield high level of privacy. In GAP, however, high distortion might dramatically reduce the classification accuracy of non-private label. This might be caused by the high correlation between private and non-private labels. This difference enables AutoGAN-DRP to preserve more utility than GAP at the same distortion level, as the experiment result (displayed in Figure 7) reveals.

In the experiment, we reproduce a prototype of Transposed Convolutional Neural Nets Privatizer (TCNNP) in GAP using materials and source code provided by [12]. We also modify our framework to make it as similar to TCNNP as possible. Specifically, a combination of two convolutional layers with ReLU activation function and two fully connected neural network layers are used for implementing the Generator similar to TCNNP. Our Classifier is constructed on two convolutional layers and two fully connected hidden layers similar to the Adversary in GAP. We also test our framework on GENKI, the same dataset with GAP. The utility is evaluated by the accuracy of facial expression classification (a binary classification). It should be noted that our framework have been shown to work on different datasets with multi-class classification, which is more challenging and comprehensive. Figure 7 shows the accuracy results of GAP and AutoGAN-DRP for GENKI dataset. AutoGAN-DRP achieves distances ranging from 0.037 to 0.039 for different dimensions from one to seven. At the same range of distance (distortion per pixel), GAP achieves accuracy of only 72% while AutoGAN-DRP gains accuracy rates starting from 77% to 91% for different number of dimensions. It becomes evident that our method can achieve higher accuracy than that of GAP at the same distortion level.

9. Conclusion

In this paper, we introduce a mathematics tool $\epsilon - DR$ to evaluate privacy preserving mechanisms. We also propose a non-linear dimension reduction framework. This framework projects data on lower dimension domain in which it prevents data reconstruction and preserve data utility. The dimensionally-reduced data can be used effectively for the machine learning tasks such as classification. Our future works plan to extend the framework for adapting with different types of data, such as time series and categorical data. We will apply different metrics to compute the distance other than l_2 norm and investigate the framework on several applications in security systems and data collaborative contributed systems.

Acknowledgment

This work is sponsored by DARPA Brandeis program under agreement number N66001-15-C-4068. The views, options, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

References

[1] R. Bost, R. Popa, S. Tu, S. Goldwasser, Machine Learning Classification over Encrypted Data, Ndss '15 (February) (2015) 1–31 (2015). doi:10.14722/ndss.2015.23241.
URL http://eprint.iacr.org/2014/331.pdf

[2] F. Emekci, O. D. Sahin, D. Agrawal, A. El Abbadi, Privacy preserving decision tree learning over multiple parties, Data Knowl. Eng. 63 (2) (2007) 348–361 (2007). doi:10.1016/j.datak.2007.02.004.

[3] E. Hesamifard, H. Takabi, M. Ghasemi, CryptoDL : Deep Neural Networks over Encrypted Data (2017) 1–21 (2017). arXiv:arXiv:1711.05189v1.

[4] A. C.-C. Yao, How to generate and exchange secrets, 27th Annu. Symp. Found. Comput. Sci. (sfcs 1986) (1) (1986) 162–167 (1986). arXiv:arXiv:1011.1669v3, doi:10.1109/SFCS.1986.25.
URL http://ieeexplore.ieee.org/document/4568207/

[5] A. Shamir, How to share a secret, Commun. ACM (1979) 612–613 (1979). doi:10.1007/978-3-642-15328-0_17.

[6] K. Chaudhuri, Privacy-preserving logistic regression 1–8.

[7] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, M. Winslett, Functional Mechanism: Regression Analysis under Differential Privacy (2012) 1364–1375 (2012). arXiv:1208.0219.
URL http://arxiv.org/abs/1208.0219

[8] N. Phan, Y. Wang, X. Wu, D. Dou, Differential Privacy Preservation for Deep Auto-Encoders: An Application of Human Behavior Prediction, Aaai (2016) 1309–1316 (2016).

[9] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep Learning with Differential Privacy (Ccs) (2016). arXiv:1607.00133, doi:10.1145/2976749.2978318.

[10] L. O.-M. Xiaoqian Jiang, Zhanglong Ji, Shuang Wang, Noman Mohammed, Samuel Cheng, Differential-Private Data Publishing Through Component Analysis, Trans. data Priv. 6 (1) (2013) 19–34 (2013). arXiv:NIHMS150003, doi:10.1080/10810730902873927.Testing.

[11] D. Zhuang, S. Wang, J. M. Chang, Fripal: Face recognition in privacy abstraction layer, in: 2017 IEEE Conference on Dependable and Secure Computing, IEEE, 2017, pp. 441–448 (2017).

[12] X. C. L. S. R. R. Chong Huang, Peter Kairouz, Generative Adversarial Privacy, Privacy in Machine Learning and Artificial Intelligence Workshop, ICML 2018 (2018).

[13] X. Chen, P. Kairouz, R. Rajagopal, Understanding compressive adversarial privacy, CoRR abs/1809.08911 (2018). arXiv:1809.08911.
URL http://arxiv.org/abs/1809.08911

[14] S. Zhou, K. Ligett, L. Wasserman, Differential privacy with compression, CoRR (2009).

[15] Z. Wu, Z. Wang, Z. Wang, H. Jin, Towards privacy-preserving visual recognition via adversarial training: A pilot study, CoRR abs/1807.08379 (2018). arXiv:1807.08379.
URL http://arxiv.org/abs/1807.08379

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Networks (2014) 1–9 (2014). arXiv:1406.2661, doi:10.1001/jamainternmed.2016.8245.
URL http://arxiv.org/abs/1406.2661

[17] P. Baldi, Autoencoders, Unsupervised Learning, and Deep Architectures, ICML Unsupervised Transf. Learn. (2012) 37–50 (2012). arXiv:1509.02971, doi:10.1561/2200000006.

[18] M. Al-rubaie, P.-y. Wu, J. M. Chang, S.-y. Kung, Privacy-Preserving PCA on Horizontally-Partitioned Data.

[19] T. Pedersen, Y. Saygn, E. Sava, Secret charing vs. encryption-based techniques for privacy preserving data mining, Fac. Eng. Nat. Sci. Sabanci Univ. Istanbul, TURKEY 1 (2007) 1–11 (2007).
URL http://research.sabanciuniv.edu/10217/

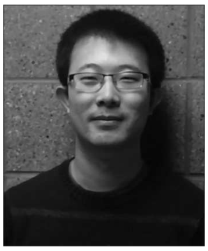
[20] C. Dwork, Differential privacy, Proc. 33rd Int. Colloq. Autom. Lang. Program. (2006) 1–12 (2006). arXiv:arXiv:1011.1669v3, doi:10.1007/11787006_1.

[21] B. Hitaj, G. Ateniese, F. Perez-Cruz, Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning (2017). arXiv:1702.07464, doi:10.1145/3133956.3134012.
URL http://arxiv.org/abs/1702.07464

[22] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: Illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intelligence 23 (6) (2001) 643–660 (2001).

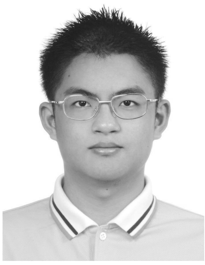
[23] F. S. Samaria, A. C. Harter, Parameterisation of a stochastic model for human face identification, in: Proceedings of 1994 IEEE Workshop on Applications of Computer Vision, 1994, pp. 138–142 (Dec 1994). doi:10.1109/ACV.1994.341300.

Hung Nguyen received the M.Sc. degree and he is currently pursuing his Ph.D. degree in Department of Electrical Engineering, University of South Florida, FL, USA. His current research interests include machine learning techniques, artificial intelligence, cyber security, privacy enhancing technologies. He is a student member of IEEE.



Di Zhuang received the B.E. degree in computer science and information security from Nankai University, China. He is currently pursuing his Ph.D. degree in electrical engineering with University of South Florida, Tampa. His research interests include cyber security, social network science, privacy enhancing technologies, machine learning and big data analytics. He is a student member of IEEE.

J. Morris Chang is a professor in the Department of Electrical Engineering at the University of South Florida. He received his Ph.D. degree from the North Carolina State University. His past industrial experiences include positions at Texas Instruments, Microelectronic Center of North Carolina and AT&T Bell Labs. He received the University Excellence in Teaching Award at Illinois Institute of Technology in 1999. His research interests include: cyber security, wireless networks, and energy efficient computer systems. In the last six years, his research projects on cyber security have been funded by DARPA. Currently, he is leading a DARPA project under Brandeis program focusing on privacy-preserving computation over Internet. He is a handling editor of Journal of Microprocessors and Microsystems and an editor of IEEE IT Professional. He is a senior member of IEEE.



PeiYuan Wu is an assistant professor at National Taiwan University since 2017. He was born in Taipei, Taiwan, R.O.C., in 1987. He received the B.S.E. degree in electrical engineering from National Taiwan University in 2009, and the M.A. and Ph.D. degrees in electrical engineering from Princeton University in 2012 and 2015, respectively. He joined Taiwan Semiconductor Manufacturing Company from 2015 to 2017. He was a recipient of the Gordon Y.S. Wu Fellowship in 2010, Outstanding Teaching Assistant Award at Princeton University in 2012. His research interest lies in artificial intelligence, signal processing, estimation and prediction, and cyber-physical system modeling.