공학석사학위논문

Privacy-Preserving Image Representations Using Sensitive Attribute Transition

민감 정보 전이를 이용한 안전한 이미지 인코딩 변환

2022년 8월

서울대학교 대학원

컴퓨터공학부

송 호 준

Privacy-Preserving Image Representations using Sensitive Attribute Transition

지도 교수 장 병 탁

이 논문을 공학석사 학위논문으로 제출함
2022년 7월

서울대학교 대학원
컴퓨터공학부
송호준

송호준의 공학석사 학위논문을 인준함
2022년 8월

위 원 장 _____ 염 헌 영 _____ (인)

부위원장 _____ 장 병 탁 _____ (인)

위 원 _____ 전 병 곤 _____ (인)

Privacy-Preserving Image Representations Using Sensitive Attribute Transition

by

HoJoon Song

A Dissertation Submitted to the Faculty of the Graduate School of Seoul National University in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

August 2022

Advisory Committee:

The manuscript printed and bound in the present form is intended for personal use and distribution among the author's colleagues, friends, and family.

An electronic copy of the document has been submitted to the Seoul National University Library with minor formatting adjustments in compliance with the University's requirements.

This dissertation contains contents of the following journal publications:
Name et al., 2021. *J. Good Sci.* 96, 105708;
Name et al., 2021. *Int. J. Nice Sci.* 31, To Appear;
List valid at the time of submission—Material from the dissertation may produce one or more additional publications in the future.

A catalogue record is available from the National Library of Korea

Typeset with LaTeX using cumg-template based on zeta709's snuthesis class
To view the author's profile, visit

Printed in South Korea
1 3 5 7 8 6 4 2

Abstract

Local Differential Privacy (LDP) is a widely accepted mathematical notion of privacy that guarantees a quantified privacy budget on sensitive data. However, it is difficult to apply LDP algorithms to unstructured data such as images since the fundamental mechanism underlying in many LDP algorithms, Randomized Response (RR), is suited for structured, tabular data. In this paper, we propose a novel task-agnostic LDP framework that preserves the privacy of selected sensitive attributes in an image representation while conserving other visual aspects. Our framework includes an adversarially trained transition model that portrays the RR mechanism, allowing it to be easily utilized in other LDP algorithms. We provide strict description of the problem formulation, and show how our model can prevent attacks from a potential adversary trying to obtain the sensitive information. Our experimental results verify that the proposed framework outperforms baseline models in protecting sensitive attributes with minimal performance loss in arbitrary downstream tasks.

Student ID: 2020-20277

Contents

Abstract **i**

Contents **ii**

List of Figures **iv**

List of Tables **v**

1 Introduction **1**

1.1 Introduction 1

2 Related Works **5**

2.1 Privacy-Preserving Machine Learning 5

2.2 Differential Privacy 6

3 Problem Formulation **9**

4 Method **11**

4.1 Attribute Inference Attack of the Adversary 11

4.1.1 Differential distance learning. 13

4.2 Task-agnostic attribute transition model 15

4.2.1 GAN Architecture. 15

4.2.2 Distributional Transition Loss 16

4.2.3 Unsensitive attribute preservation 19

4.3 Local Differentially Private Image Representation Transition Framework 20

5 Experimental Results **21**

5.1 Experimental Setup 21

5.2 Multi-Label Classification 21

5.2.1 Sensitive attribute transition evaluation 23

5.3 Evaluation on Other Attributes 26

5.4 Qualitative Results 28

5.5 Experiment on CheXpert dataset 31

6 Conclusion **32**

Bibliography **33**

초 록 **39**

List of Figures

1.1 An adversary who can perform black-box attacks (c) can easily perform attribute inference attacks on data, even if the data was encoded by the data provider (a). Our model performs privacy-preserving transitions (b) to deceive the adversary. 2

4.1 We use a modified unet architecture, (b) as the generator for our model. By substituting the concatenation-convolution layers with a summation layer at the end, our model learns the differences between representations that decide the sensitive attribute. 12

4.2 Overall architecture of the proposed transition model. The classifier in (c) is first pretrained by the same way as the adversary's training method as shown in (e). Then the classifier is frozen and used for training the generator. 14

4.3 A diagram showing the resulting distributions made by different losses. (b) shows that the cross entropy loss conflicts with the GAN loss and generates a distribution different from the original distribution like (c). (d) shows that using the distributional transition loss can generate distributions similar to the original. 17

5.1 Examples of data used in our experiments. (a) CelebA, (b) CheXpert 22

5.2 Qualitative results of our transition model 29

List of Tables

5.1 Results of sensitive attribute transition evaluated using different baselines and our model 25

5.2 Results of unsensitive attribute correlation preservation where the sensitive attribute is $Male$ and $\epsilon = 1$. 27

5.3 Results of experiment on CheXpert dataset with Edema as the sensitive attribute. The abbreviations refer respectively, Cardiomegaly, Enlarged Cardiomediastinum, Consolidation, and Pleural Effusion. 30

1 Introduction

1.1 Introduction

While many traditional data analysis methods have struggled with handling unstructured, non-tabular data, recent studies in deep learning have shown promising performances in analyzing visual data(Dosovitskiy et al. 2020; Goodfellow et al. 2014; Radford et al. 2021; Russakovsky et al. 2015). However, powerful analysis methods aligned with the data-driven nature of deep learning bring up new privacy concerns: *How do we protect private information in images*? Protecting the sensitive attributes in visual data without affecting other features is a challenging problem since every semantic visual attributes and features are intimately entangled together into a single image.

In this paper, we address the scenario where a data provider wants to share their data with a data analyst, but cannot fully trust the analyst (or the data transmission process itself) as their data may include some highly sensitive information. This is a common case in various industries, for example, a hospital (i.e. the data provider) may request an ML company (the data analyst) a model that can diagnose a specific disease by analyzing various examination results. The hospital does not want the service provider or potential adversaries to be able to extract any sensitive information that may identify or infringe the privacy of their patients.

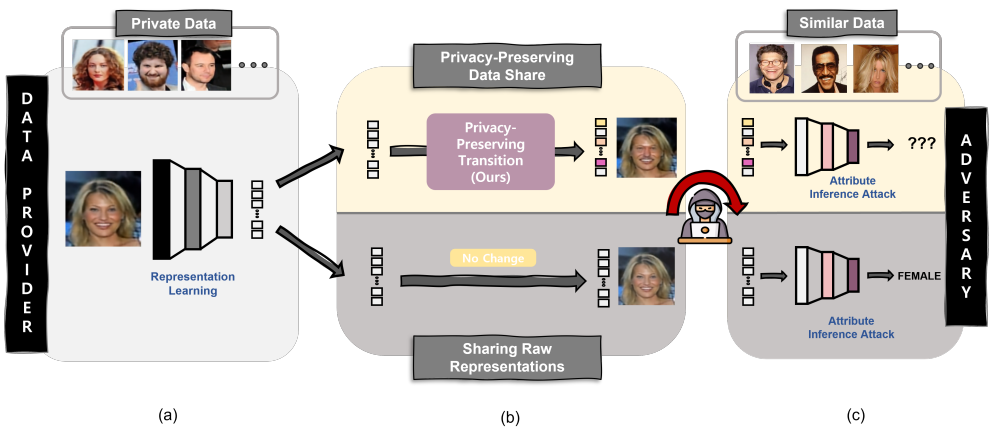


Figure 1.1 An adversary who can perform black-box attacks (c) can easily perform attribute inference attacks on data, even if the data was encoded by the data provider (a). Our model performs privacy-preserving transitions (b) to deceive the adversary.

Providing encoded representations instead of raw data is a simple approach to keep adversaries from acquiring the original data, but isn't a fundamental solution considering that the adversaries' objective is focused on extracting the sensitive attributes instead of the whole data(Ganju et al. 2018).

Many previous works have tackled privacy-preserving in visual data by using adversarial training methods to create representations that preserve the features useful for the target task while sanitizing the private features(Edwards and Storkey 2016; Xiao et al. 2020; Pittaluga et al. 2019; Xiong et al. 2019; Chen et al. 2018). Although these approaches have shown promising results, there remains several potential problems: First, they can only be applied in cases where the target task is known to the data provider in advance. Second, the target downstream task is assumed to be independent of the protected attributes. Third, only one sensitive attribute can be protected while obtaining the representation.

This paper proposes a novel framework that addresses the aforementioned problems by applying adversarial training to representations to portray a *randomized response*(Warner 1965; Mangat 1994) (RR) mechanism to achieve *local differential privacy*(Dwork and Roth 2013; Kasiviswanathan et al. 2011) (LDP), a rigorous mathematical definition of privacy that quantifies the privacy budget of privacy-preserving algorithms. Randomized response is a traditional, but powerful method used to achieve local differential privacy by substituting a local entity randomly with another value. The randomized process guarantees plausible deniability with respect to the target attribute domain for each individual in the data.

We show experimental evidence that our framework can generate new representations that change the private attributes while conserving other

independent attributes and inherent correlations without any training related to potential target tasks.

Our contributions are summarized as follows.

- We suggest a novel task-agnostic adversarial training model that learns the difference vector between representations that differ in only the sensitive attribute.
- We propose a local differentially private protocol that utilizes our new model to portray the random response mechanism used in existing LDP methods. To the extent of our knowledge, we are the first to propose a framework that quantifies the trade-off between utility and the privacy budget of the sensitive attributes in images.
- We evaluate our methods on two multi-label classification datasets of different domain, CelebA(Liu et al. 2015, 2018) and CheXpert(Irvin et al. 2019), to show that our method can be applied to arbitrary domains and attributes.
- We provide mathematical analysis of the results showing that our framework maintains the statistical properties of the data, including the inherent correlation between features.

2 Related Works

2.1 Privacy-Preserving Machine Learning

Many studies have proved that sensitive information could be unintentionally leaked while utilizing deep learning models due to the data-driven nature of deep neural networks(Nasr et al. 2019; Gong and Liu 2016; Fredrikson et al. 2015; Carlini et al. 2021). The interest for preserving privacy and defense methods for such leakages and attacks in machine learning is growing(Melis et al. 2019; Ying et al. 2020; Hardt et al. 2012).

A popular approach for privacy-preserving methods is using deep models that utilize adversarial training to sanitize the sensitive attribute from the data. The adversary is trained to infer the sensitive attribute from the model's outputs while the model is trained to generate outputs which make the adversary's inference fail. (Edwards and Storkey 2016) showed that adversarial training could be used to remove certain attributes in images, (Xiong et al. 2019) showed that a similar model could be applied to data collected during auto-driving, and (Ren et al. 2018) used a similar approach to action recognition tasks while protecting individual identities.

2.2 Differential Privacy

Differential privacy(Dwork 2008) is a rigorous definition of privacy that guarantees a quantified boundary (the privacy budget ϵ) of an individual's privacy leakage possible to a potential adversary. The intuition is that there exists some probability of a mechanism giving the same output on two datasets that differentiate by one individual record. This means that the presence of an individual in the dataset cannot be distinguished by the mechanism. For the formal definition of differential privacy, we refer to (Dwork and Roth 2013):

Definition 2.1 ($\boldsymbol{\epsilon}$-Differential Privacy) A randomized algorithm $\mathcal{M}$ is ϵ-differential private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for every adjacent datasets $\mathcal{D}, \mathcal{D}'$, the following equation holds:

$$Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon} Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] \tag{2.1}$$

One drawback of differential privacy is that there must exist a trusted data aggregator, who makes differentially private computations on the sensitive data. In the case where the data provider cannot trust the data aggregator, local differential privacy(Kasiviswanathan et al. 2011) can be used. LDP is achieved when the data provider applies a differentially private mechanism to its data before transferring it to the data aggregator. The aggregator cannot do anything to break the privacy already achieved by the LDP mechanism since differentially private mechanisms are closed under post-processing. The formal description of local differential privacy(Erlingsson et al. 2014) is as follows:

Definition 2.2 ($\boldsymbol{\epsilon}$-Local Differential Privacy) A randomized algorithm A is ϵ-local differential private if for all $\mathcal{O} \subseteq \text{Range}(A)$ and for every possible pairs

of data inputs v, v', the following equation holds:

$$Pr[A(v) \in \mathcal{O}] \leq e^{\epsilon} Pr[A(v') \in \mathcal{O}] \tag{2.2}$$

Differential privacy algorithms can provide a stronger, quantified degree of privacy compared to deanonymizing or sanitization techniques as these techniques were revealed to be vulnerable to linkage attacks(Narayanan and Shmatikov 2006, 2008; Kosinski et al. 2013; El Emam et al. 2011). While the common practice for differential privacy is to inject random errors based on the privacy budget, most LDP mechanisms are built on top of the Randomized Response (RR) technique.

The intuition of RR is that if an individual answers a sensitive question with a random probability, the individual can gain plausible deniability to the recorded answer. For example, consider the case when an individual participating in a survey is asked "Are you a smoker?" and is told to answer with probability $(1-p)$ the correct answer, and with probability p a random answer of yes/no. There is no way for the surveyor to find out if the respondent's answer is based on the true experience or the randomness of the mechanism. The surveyor can still gain statistical utilities (in this case, the fraction of smoking people in the surveyed population) depending on p. A low p will give more accurate statistics but less privacy for individuals, and a high p will guarantee higher privacy but inaccurate results. The effectiveness of RR-based LDP mechanisms are widely utilized in production to ensure the local privacy of individuals(Erlingsson et al. 2014; Differential Privacy Team 2017; Cormode et al. 2018; Ding et al. 2017).

A potential weakness of RR is that it is difficult to be applied on non-structured, non-tabular data such as images. The methodology proposed in

this paper attempts to address this issue by considering image representations as sets of semantic visual attributes entangled together. We simulate a RR mechanism with respect to sensitive attributes by training a model to learn the differences of two representations that differ in only the target attribute. This can be thought similar to changing a single column of a row in a tabular dataset.

3 Problem Formulation

Throughout this section and the rest of this paper, we refer to privacy preserving taxonomy organized in (Rigaki and Garcia 2020). Our main objective is to create a safe framework for a data provider who wants to release or share their data containing sensitive attributes. The 'sensitive attributes' are assumed to be discrete, categorical variables (e.g. the gender of a person). For simplicity, we consider only binary domains in this paper but our method can be applied to k-ary domains without variation, by repeating the method on each domain attribute. We refer to other attributes inherent in the representation that are independent of the sensitive attribute as unsensitive attributes.

The data provider decides to share their data in an encoded form for various reasons (e.g. applying other privacy-preserving encoding methods, changing dimensionality of the data, reducing data size, etc.), which is a common practice(Abu-El-Haija et al. 2016; Cazzolato et al. 2021). But the problem is that there exist potential adversaries trying to find out the sensitive information from individual data records. We assume that the adversaries are under the following settings, similar to the adversary proposed in (Xiao et al. 2020):

- Adversaries perform black-box attacks, meaning they can make arbitrary inferences to the encoder model.

- They don't have access to the original data, but have access to a dataset similar to the original data. By 'similar', we mean they are drawn from the same distribution but differ in individual records.
- Their goal is to perform inference attacks on the sensitive attributes inherent in the shared representations.

We assume the data provider is unaware of downstream tasks to be performed on the representations, meaning our protection scheme cannot target achieving satisfactory performance of specific tasks or preserving selected attributes like the settings in previous works. There may be multiple sensitive attributes that need protection, as many studies proved that protecting unique identifiers are highly insufficient to prevent de-anonymization attacks(Narayanan et al. 2011; Gambs et al. 2014; Narayanan and Shmatikov 2008).

Our framework probabilistically substitutes the private attributes which exist implicitly in the representations, so that the substituted representations deceives the adversaries' attack models that are successful to the original representations.

4 Method

4.1 Attribute Inference Attack of the Adversary

The presumed adversaries can easily train inference models that extract the sensitive attributes by creating input-representation pairs with their own data(Xiao et al. 2020). With $\mathcal{D}_{orig}$ as the original dataset, an adversary has access to the encoder E, and its own adversarial dataset $\mathcal{D}_{adv}$, which is drawn from the same distribution as $\mathcal{D}_{orig}$. The objective of the adversary is to build a classifier $C_s : \mathcal{Z} \rightarrow \mathcal{S}$ where $\mathcal{Z} = \left\{E(X) \mid X \in \mathcal{D}_{orig}\right\}$ and $\mathcal{S}$ is the range of the sensitive attribute s (which is $[0, 1]$, as we limited our problem to the binary domain). Since $\mathcal{S}$ is a set of discrete categories, the adversary can easily build C_s by using the cross entropy loss:

$$\mathcal{L}_{adv} = \mathbb{E}_{X \in \mathcal{D}} \left[X_s \cdot \log \hat{X}_s \right] \tag{4.1}$$

where X_s is the true label of the private attribute of X and $\hat{X}_s$ is the predicted label from $C_s(E(X))$. Note that since $\mathcal{D}_{orig}$ and $\mathcal{D}_{adv}$ share the same distribution, a classifier trained using $\mathcal{D}_{adv}$ will show similar performance on $\mathcal{D}_{orig}$ as long as $\mathcal{D}_{adv}$ has enough entries to generalize.

In order to protect the private attribute while not losing information that may be used for potential downstream tasks, we train the main transition module of our proposed based on two objectives : 1) the model must change

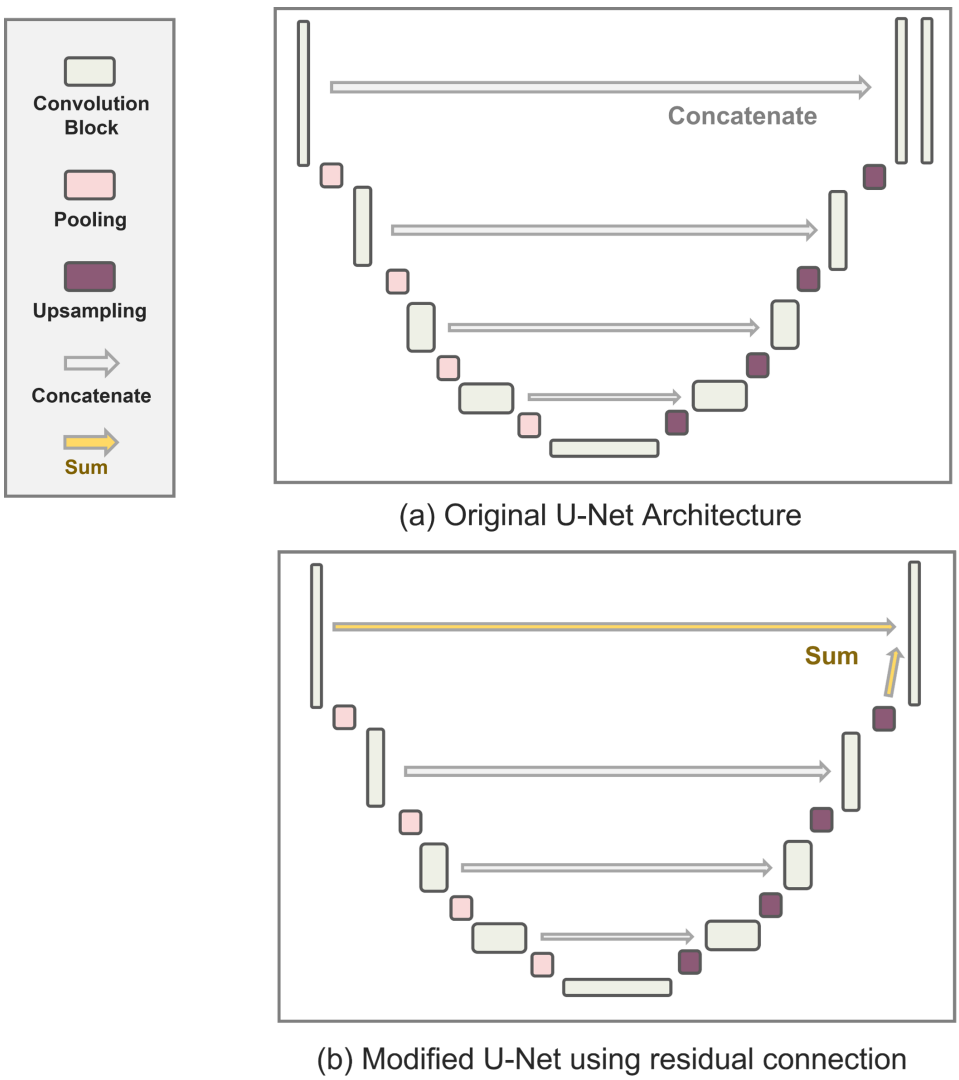


Figure 4.1 We use a modified unet architecture, (b) as the generator for our model. By substituting the concatenation-convolution layers with a summation layer at the end, our model learns the differences between representations that decide the sensitive attribute.

the implicit private attribute and 2) the model must preserve other unsensitive attributes.

The two objectives share a similar goal to image-to-image translation tasks(Isola et al. 2017; Kim et al. 2017; Choi et al. 2018; Zhu et al. 2017). Image translation models aim to transform images into another domain while maintaining the semantic visual features of the original image. However, the sensitive attribute transition task and image-to-image translation task differ in terms of spatial locality. Whilst image-to-image translation focuses on changing the global spatial distribution of the entire image (e.g. changing photos to drawings, images to semantic segmentation labels, etc.), our objective tends to focus on changing local attributes from a latent representation (e.g. gender of a person, indicators of a specific disease in an x-ray image, etc.).

4.1.1 Differential distance learning.

Many image translation models use an encoder-decoder based generation model such as U-net(Ronneberger et al. 2015) and Resnet(He et al. 2016)-based generators to generate images that fit the distribution of the target domain with the features extracted by the encoder as input. However, our transition task focuses on changing local attributes instead of reconstructing the features into a new distribution. Using the same encoder-decoder architecture for our task will cause training overhead of having to learn identity functions for unsensitive attributes. We apply a simple idea of learning the difference vector between two representations that differ in only the attribute to address the training overhead.

Therefore, the objective of our attribute transition model is to build a generator $G_{s_1 s_2}$ which takes a representation Z as input and outputs Z' where

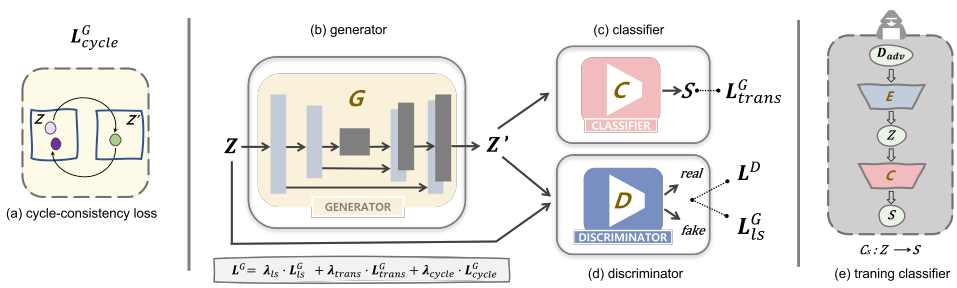


Figure 4.2 Overall architecture of the proposed transition model. The classifier in (c) is first pretrained by the same way as the adversary's training method as shown in (e). Then the classifier is frozen and used for training the generator.

$\Delta = Z' - Z$ given $s_1, s_2 \in \mathcal{S}, Z_s = s_1, Z'_s = s_2$ and all unsensitive attributes for Z, Z' are equal. Instead of directly learning Z', we train $G_{s_1 s_2}$ to learn Δ, which shares the same dimensionality as Z and Z'. The idea of learning the difference vector can be easily understood as a model-level residual skip connection from the resnet(He et al. 2016) architecture. As U-net generators already utilize long skip connections with channel-wise concatenation, our model uses a U-net generator with the final concatenation-convolution layer substituted with element-wise summation as shown in Figure 4.1.

4.2 Task-agnostic attribute transition model

In this section, we describe the loss functions we use to train our transition model. The overall architecture of the model is shown in Figure 4.2.

4.2.1 GAN Architecture.

We use the Generative Adversarial Training(GAN)(Goodfellow et al. 2014) architecture to train our model. Generators in GANs attempt to create new data entries that look realistic enough to fool the discriminator, while the discriminator is trained to distinguish original data and generated data. We use the least square loss of LSGAN proposed by (Mao et al. 2017) as the GAN loss. The loss function for the discriminator D optimizes its ability to label real representations as real and generated representations as fake:

$$\mathcal{L}^D = \frac{1}{2}\mathbb{E}_{Z \in \mathcal{Z}}[(D(Z) - a)^2] + \frac{1}{2}\mathbb{E}_{Z \in \mathcal{Z}}[(D(G(Z)) - b)^2] \tag{4.2}$$

where a and b are the labels for the real representations and fake representations respectively. The loss function for the generator $G_{s_1 s_2}$ optimizes its ability to

fool D, making it label the generated representations as real.

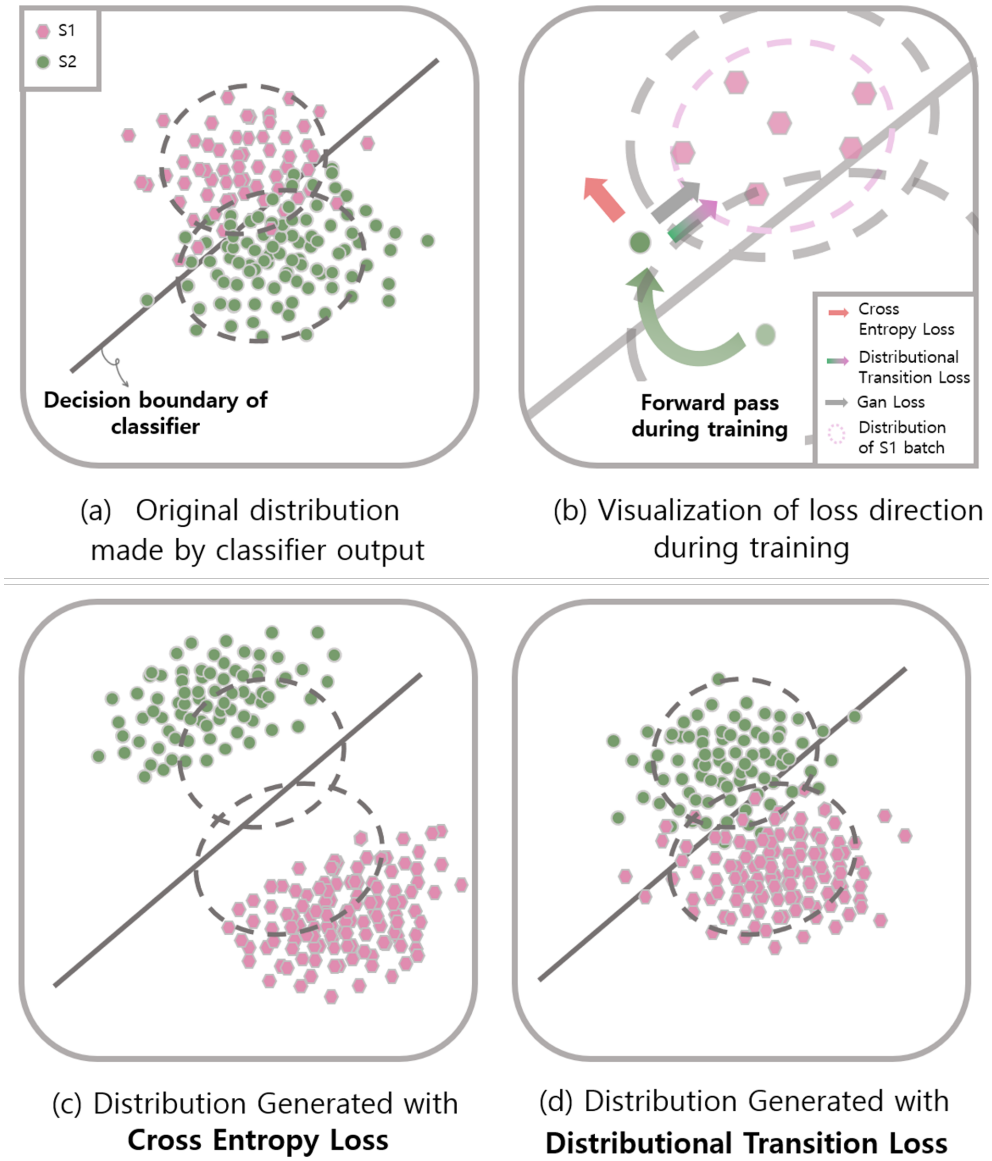
$$\mathcal{L}_{ls}^{G_{s_1 s_2}} = \mathbb{E}_{Z \in \mathcal{Z}}[(D(G_{s_1 s_2}(Z)) - a)^2] \tag{4.3}$$

4.2.2 Distributional Transition Loss

The main objective of the transition model is to fool the adversary into obtaining a false label in an attempt to infer the private attribute from the representation. Our presumed adversaries are under powerful settings of which they can use their own similar dataset to build a classifier that can identify sensitive attributes of representations. We directly use the adversary's classifier with frozen weights as an auxiliary classifier to change the sensitive attribute in generator outputs.

Recall that our transition model uses pre-encoded representations made from arbitrary encoders[1] designed for various purposes (e.g. obtaining disentangled representations, other privacy-preserving methods, data compression, etc.) Using latent representations made by arbitrary encoders implies that the attributes implicit in the original data may be compressed, partially lost, or transformed in the encoding process. Such data loss and the fundamental limitations of deep classifiers may lead to unclear decision boundaries for the adversary's classifier as shown in Figure 4.3 (a). Therefore, maximizing the likelihood of the classifier's output for the generator (like many GANs that utilize auxiliary encoders(Odena et al. 2017; Choi et al. 2018)) using the cross entropy loss may lead to generating results distant from the original data's distribution. The cross entropy loss will conflict with the GAN loss as shown in Figure 4.3 (b), causing the training to diverge and generate low-quality results.

[1]Using no encoders can also be an option, considering that raw data are identical representations of themselves.



(a) Original distribution made by classifier output

(b) Visualization of loss direction during training

(c) Distribution Generated with **Cross Entropy Loss**

(d) Distribution Generated with **Distributional Transition Loss**

Figure 4.3 A diagram showing the resulting distributions made by different losses. (b) shows that the cross entropy loss conflicts with the GAN loss and generates a distribution different from the original distribution like (c). (d) shows that using the distributional transition loss can generate distributions similar to the original.

We propose to minimize the KL-divergence between the distribution made by classifier results of the original representations where $Z_s = s_2$ and the distribution made by classifier results of $G_{s_1 s_2}(Z)$, where $Z_s = s_1$ in order to deceive the adversary's classifier while not conflicting with the GAN loss. This also satisfies our objective of preserving statistical properties of the original representations, as we can consider well-trained classifiers as a metric function for categorical attributes(Lei 2014).

With $P(x \mid \theta)$ as the distribution of the classifier outputs of the original representations and $P(x \mid \theta^*)$ as the distribution of classifier outputs of the generated representations, we prove that minimizing the KL-divergence for $P(x \mid \theta)$ and $P(x \mid \theta^*)$ is equal to minimizing the negative log-likelihood of the classifier outputs of generated representations to be observed under $P(x \mid \theta)$.

$$\begin{aligned}
\underset{\theta}{\operatorname{argmin}}\ D_{KL}[P(x \mid \theta^*) \| P(x \mid \theta)] &= \underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{x \sim P(x|\theta^*)}[\log \frac{P(x \mid \theta^*)}{P(x \mid \theta)}] \\
&= \underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{x \sim P(x|\theta^*)}[\log P(x \mid \theta^*) - \log P(x \mid \theta)] \\
&= \underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{x \sim P(x|\theta^*)}[-\log P(x \mid \theta)] \\
&= \underset{\theta}{\operatorname{argmin}}\ -\frac{1}{n}\sum_{i}^{n} \log P(x_i \mid \theta)
\end{aligned} \tag{4.4}$$

where $x_1, x_2, \ldots, x_n$ are classifier outputs for generated representations in a batch and the last term holds by the law of large numbers (we assume a large batch size). For mathematical convenience, we assume a normal distribution for $P(x \mid \theta) = \mathcal{N}(\mu, \sigma^2)$ where μ and σ^2 are the mean and variance of $\{C_s(Z) \mid Z \in \mathcal{Z}\}$. The negative log-likelihood of given observations under a normal distribution can be derived as:

$$
\begin{aligned}
l(\mu, \sigma^2; x_1, x_2, ..., x_n) &= \log\left[(2\pi\sigma^2)^{-n/2} exp(-\frac{1}{2\sigma^2}\sum_i^n (x_i - \mu)^2)\right] \\
&= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i^n (x_i - \mu)^2
\end{aligned} \tag{4.5}
$$

As the first two terms are uneffected by θ, we formulate our loss function for sensitive attribute transition as follows:

$$
\mathcal{L}^G_{trans} = \mathbb{E}_{Z \in \mathcal{Z}_{s_1}}[(C_s(G_{s_1 s_2}(Z)) - \mu)^2] \tag{4.6}
$$

where $\mu = \mathbb{E}_{Z \in \mathcal{Z}_{s_2}}[C_s(Z)]$ and $\mathcal{Z}_{s_i} = \{E(X) \mid X \in \mathcal{D}_{orig}, X_s = s_i\}$.

4.2.3 Unsensitive attribute preservation

To preserve the attributes that are independent of the sensitive attribute, we apply the cycle-consistency loss proposed in (Zhu et al. 2017).

$$
\mathcal{L}^G_{cycle} = \frac{1}{2}(\mathbb{E}_{Z \in \mathcal{Z}_{s_1}}[\|G_{s_2 s_1}(G_{s_1 s_2}(Z)) - Z\|_1] + \mathbb{E}_{Z \in \mathcal{Z}_{s_2}}[\|G_{s_1 s_2}(G_{s_2 s_1}(Z)) - Z\|_1) \tag{4.7}
$$

The cycle consistency loss allows the generator to preserve the unsensitive attributes since any modifications to an unsensitive attribute by $G_{s_1 s_2}$ may not be reverted by $G_{s_2 s_1}$ because it is independent of s.

Overall, we optimize generator $G_{s_1 s_2}$ and $G_{s_2 s_1}$ jointly with the weighed sum of the proposed losses:

$$
\mathcal{L}^G = \lambda_{ls} \cdot \mathcal{L}^G_{ls} + \lambda_{trans} \cdot \mathcal{L}^G_{trans} + \lambda_{cycle} \cdot \mathcal{L}^G_{cycle} \tag{4.8}
$$

4.3 Local Differentially Private Image Representation Transition Framework

The proposed transition model allows us to treat an image representation as a single row in structured tabular data, with semantic visual attributes as columns. The model learns to change a single column (the targeted sensitive attribute) in the image, without collapsing the structure of the data or affecting columns independent of the sensitive attribute. Attributes dependent or correlated to the sensitive attribute will also be perturbed along with the sensitive attribute depending on their correlation. We use the proposed model to portray a RR mechanism with respect to the target attribute to achieve LDP. As any RR-based LDP mechanism can be used with our model, we demonstrate our framework using the most basic and oldest form of RR proposed by Warner et al(Warner 1965):

$$\mathcal{Z}_{safe} = \left\{ Z' \mid X \in \mathcal{D} \right\},$$
$$Z' = \begin{cases} Z & \text{with probability } \frac{e^{\epsilon}}{e^{\epsilon}+1} \\ G(Z) & \text{with probability } \frac{1}{e^{\epsilon}+1} \end{cases} \tag{4.9}$$

where $Z = E(X)$ and G refers to the corresponding transition model depending on X_s. The newly formed dataset $\mathcal{Z}_{safe}$ is ϵ-LDP with respect to the sensitive attribute s.

5 Experimental Results

5.1 Experimental Setup

Our framework allows an arbitrary choice of encoders, freely chosen by the data provider depending on their objective. Since we assume no prior about the data provider, we demonstrate our method using the encoder from a trained vanilla autoencoder. Although it lacks semantic meaningfulness, the latent representation made from a vanilla autoencoder contains detailed information about the original data with minimum information loss.

We use two multi-label classification datasets, CelebA(Liu et al. 2015) and CheXpert(Irvin et al. 2019). CelebA is a large-scale face dataset containing more than 200k images of celebrities with 40 different labeled binary attributes. We use the aligned version where the faces are cropped and aligned to the center. CheXpert is a medical dataset consisting of over 220k chest radiographs from 65k patients, including labels for 14 different observations in the radiographs.

5.2 Multi-Label Classification

We perform quantitative analysis based on a multi-label classification task on the translated representations made by our framework. Classifying each attribute in multi-label classification can be considered as arbitrary downstream

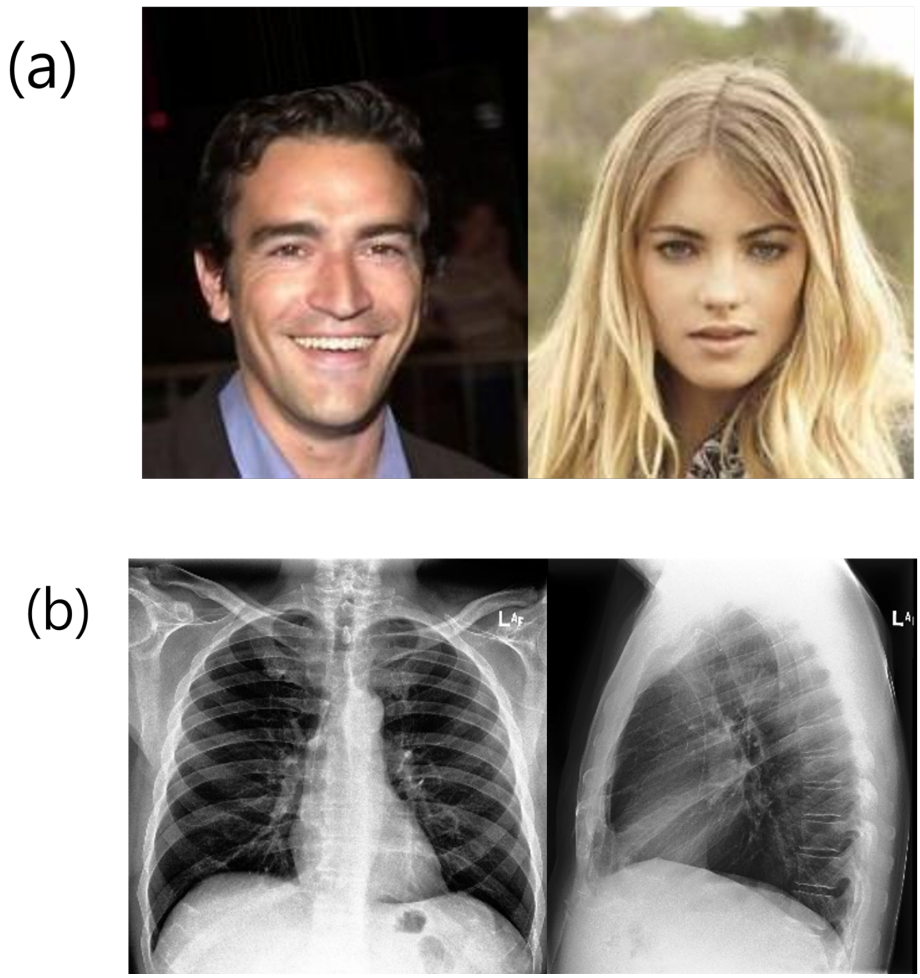


Figure 5.1 Examples of data used in our experiments. (a) CelebA, (b) CheXpert

tasks since different attributes are distinguished by different semantic visual features in the image.

Like the adversary's classifier used for training, a classifier is trained independently for each attribute on the original representations. We analyze the classification performance for each attribute differently depending on the attribute's correlation with the private attribute. Note that the objective of our experiments is not achieving high classification performance, but proving that our generated representations show consistent performance on pretrained classifiers before and after the transition. The classifiers are trained only once using the original representations and labels, and not retrained again with the generated representations, which allows them to be used as consistent metric functions for successful transition.

5.2.1 Sensitive attribute transition evaluation

For a given attribute a_i, let C_{a_i} denote the classifier trained using the cross entropy loss on the original representations $\mathcal{Z}$ made from the training set. With a_s as the sensitive attribute, let $M_{a_s} = \begin{bmatrix} p_{\text{TN}} & p_{\text{FP}} \\ p_{\text{FN}} & p_{\text{TP}} \end{bmatrix}$ denote the confusion matrix normalized so the sum of each row elements is 1, obtained by evaluating C_{a_s} on the original representations made from the validation set. Let's denote p_{s_0}, p_{s_1} as the fraction of data entries in the validation set with the sensitive label of 0 and 1 respectively. Assume an ideal transition that outputs representations with opposite sensitive attributes while maintaining the distributional properties. The classifier will maintain its performance since the new representations form a distribution nearly identical to the original distribution. If we apply the ideal transition with the RR mechanism in Equation (11) to the validation set and

evaluate the transitioned representations using C_{a_s}, we can calculate a new confusion matrix:

$$M_{a_s}{}' = \begin{bmatrix} \frac{e^{\epsilon}}{e^{\epsilon}+1} \cdot p_{s_0} & \frac{1}{e^{\epsilon}+1} \cdot p_{s_0} \\ \frac{1}{e^{\epsilon}+1} \cdot p_{s_1} & \frac{e^{\epsilon}}{e^{\epsilon}+1} \cdot p_{s_1} \end{bmatrix} \times \begin{bmatrix} p_{\text{TN}} & p_{\text{FP}} \\ p_{\text{FN}} & p_{\text{TP}} \end{bmatrix} \tag{5.1}$$

We calculate the accuracy and f1 scores of C_{a_s} on representations generated by our transition model with the RR mechanism and compare it with the ideal scores calculated by $M_{a_s}{}'$. Performance close to the ideal performance means that the model has successfully learned a transition that outputs a distribution similar with the original data. We compare the performance of our model with popular image-to-image translation models as baseline. Since the baseline models are best fit for images as inputs, we train and apply transitions to the images and then encode them instead of training them to create transitions for the encoded representations. This is a highly generous setting for baseline models since they can learn more specific mappings and spatial relationships concerned with the sensitive attribute that may have been lost in the encoding process.

Table 5.1 summarizes the results for analysis on sensitive attribute transition on the CelebA dataset. Our model outperforms the baselines for different values of ϵ on two different attributes, 'Male' and 'Smiling'. Our model shows performance close to the ideal transition, regardless of the privacy budget ϵ. Smaller ϵ means higher privacy, and when $\epsilon = 0$ privacy is fully protected but utility is lost.

Table 5.1 Results of sensitive attribute transition evaluated using different baselines and our model

	$attr =$ **Male**					
	$\epsilon = 0$		$\epsilon = 0.5$		$\epsilon = 1$	
	$f1$	acc	$f1$	acc	$f1$	acc
Original	.971	.975	.971	.975	.971	.975
Ideal	.461	.500	.578	.616	.686	.719
CycleGAN	.751	.790	.807	.836	.850	.872
StarGAN	.527	.566	.624	.662	.721	.752
Ours	**.457**	**.497**	**.580**	**.616**	**.685**	**.718**
	$attr =$ **Smiling**					
	$\epsilon = 0$		$\epsilon = 0.5$		$\epsilon = 1$	
	$f1$	acc	$f1$	acc	$f1$	acc
Original	.913	.916	.913	.916	.913	.916
Ideal	.491	.500	.593	.602	.684	.692
CycleGAN	.795	.798	.822	.826	.850	.854
StarGAN	.603	.616	.679	.690	.750	.758
Ours	**.480**	**.490**	**.587**	**.594**	**.675**	**.684**

5.3 Evaluation on Other Attributes

Another important objective of the proposed framework is preserving unsensitive attributes. But before measuring the preservation of an attribute, we must first assess its correlation with the sensitive attribute. There are two reasons that an ideal sensitive attribute transition model should maintain the inherent correlation between the attributes in an image.

First, the underlying correlations between attributes are valuable statistical properties. Second, correlated attributes can be used in linkage attacks. For example, consider the attributes 'Male' and 'Wearing Lipstick' in the CelebA dataset. 99.4% of people who are wearing lipsticks are females and 78.5% of people who aren't are males in the CelebA dataset. Even if a transition model successfully perturbed the gender attribute, the adversary can perform a linkage attack on the generated representation by inferring the 'Wearing Lipstick' attribute. Therefore, an ideal transition model should also perturb attributes highly correlated with the target sensitive attribute.

We divide the remaining attributes into two groups based on their correlation calculated using $Theil's\ U$((Theil 1971)), a widely accepted asymmetric correlation measure for categorical data, with 0.1 as the threshold. $Theil's\ U$ can be easily interpreted as the mathematical probability of a linkage attack on attribute x using the value of y. For correlated attributes, we compare the $Theil's\ U$ values $U(X|Y)$ calculated by classifier outputs before and after the transition where X is the sensitive attribute and Y is the correlated attribute. For uncorrelated attributes, we compare the classifiers' performance before and after the transition to test if they are preserved.

Table 5.2 Results of unsensitive attribute correlation preservation where the sensitive attribute is $Male$ and $\epsilon = 1$.

	Correlated Attr. (Theil's U)				
	Beard	Lipstick	Makeup	Blonde	
Ideal	0.25	0.65	0.46	0.12	
CycleGAN	**0.19**	0.46	0.3	**.09**	
StarGAN	0.18	0.49	0.36	0.05	
Ours	0.17	**0.55**	**0.38**	0.04	
	Independent Attr. (f1)				
	M.Open	Smiling	Glasses	Young	Chubby
Ideal	.893	.913	.863	.906	.418
CycleGAN	.885	**.909**	**.862**	**.899**	.364
StarGAN	.881	.934	.853	.895	**.377**
Ours	**.889**	.903	.843	.889	.359

The results for evaluation on unsensitive attributes is summarized in Table 5.2. Our model shows best performance for maintaining highly correlated attributes and similar performance to the best performance for slightly less-correlated attributes. For independent attributes, the baseline models and our model all show similar performances. This is because all three models use the cycle-consistency loss for maintaining independent attributes.

5.4 Qualitative Results

We present qualitative results of our transition model at Figure 5.2. The series of three images for each setting shows the original image, translated image, and a heatmap image drawn based on their pixelwise differences of the first and second images. Note that the images are decoder outputs of the original/generated representations as our representations were encoded by an autoencoder. The decoders of vanilla autoencoders aren't normally suitable for generation tasks because the latent space of the representations is sparse, thus making interpolation between latent representations difficult. Despite such difficulties, our qualitative results show natural-looking images, meaning that the model generates representations that belong to the sparse latent space of original representations.

The heatmap images show that the model successfully learned to target the local differences between representations that differ by only the sensitive attribute. We can see in the heatmap that only local changes related to the sensitive attribute (e.g. near the eyes for glasses, the mouth for smiling, etc.) were applied during the transition. Also, the third row shows that our transition model works soundly even if multiple target attributes are perturbed.

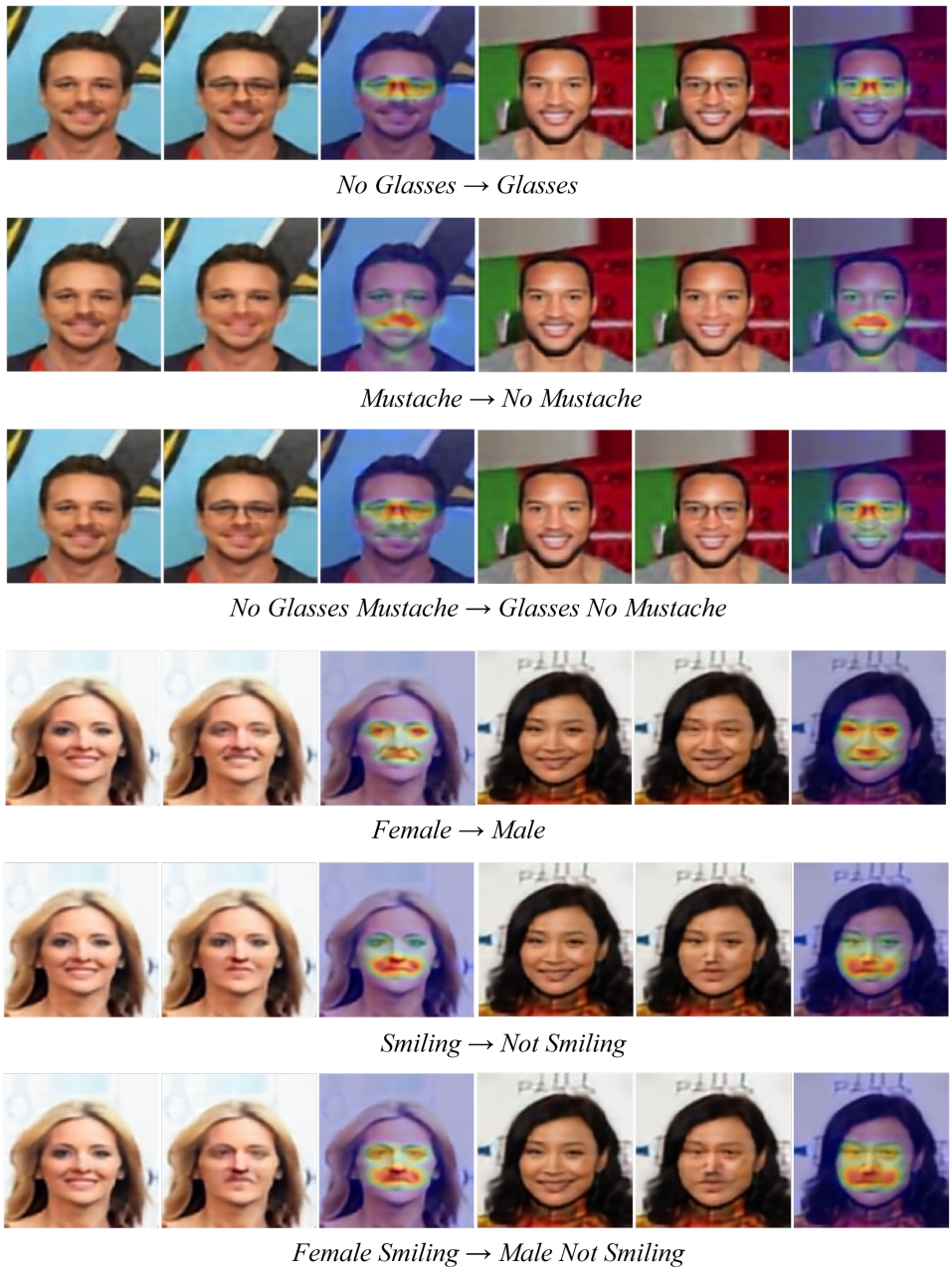


Figure 5.2 Qualitative results of our transition model

Table 5.3 Results of experiment on CheXpert dataset with Edema as the sensitive attribute. The abbreviations refer respectively, Cardiomegaly, Enlarged Cardiomediastinum, Consolidation, and Pleural Effusion.

	Edema(*sensitive*)					
ϵ	f1	acc	C.m.	E.C.	Consol.	P.E.
∞	.240	.812	.769	.877	.881	.806
0	.125(.113)	.792(.746)	.758	.866	.876	.806
.5	.157(.141)	.787(.763)	.764	.870	.879	.804
1	.192(.167)	.792(.777)	.765	.865	.872	.803

5.5 Experiment on CheXpert dataset

We provide supplementary experimental results of our model on the CheXpert dataset. Table 5.3 summarizes the results, with diagnosis for 'Edema' as the sensitive attribute. $\epsilon = \infty$ refers to the case when original representations as no privacy is preserved. The values for the unsensitive attributes(which all of them are independent with Edema) refer to the AUROC score of the classifier, often used for imbalanced data classification. The value inside the parenthesis refers to the ideal score calculated as in section 4.2. The experimental results for the CheXpert dataset prove that our framework functions on images of various domains.

6 Conclusion

In this paper, we present a novel framework that achieves ϵ-local different privacy with respect to private attributes in images. To the knowledge of our extent, we are the first to propose a probabilistic framework that controls the trade-off between privacy and utility of the data regarding the private attributes in an image. The main module in our framework is the attribute transition model which is carefully trained to deceive an attribute inference attack made by a powerful adversary. We provide theoretical reasoning and detailed description on how our loss functions are formulated. The attribute transition model also preserves the inherent statistical properties in an image and semantic visual features that are independent of the sensitive attribute. Because of such properties, our model can portray a randomized response mechanism on images, allowing it to be utilized in other LDP algorithms. We provide both quantitative and qualitative experiments that prove our claims, showing best performance compared to the baseline models.

Bibliography

Abu-El-Haija, Sami, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan, 2016: Youtube-8m: a large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al., 2021: Extracting training data from large language models. *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650.

Cazzolato, Mirela T, Lucas C Scabora, Guilherme F Zabot, Marco A Gutierrez, Caetano Traina Jr, and Agma JM Traina, 2021: Featset: a compilation of visual features extracted from public image datasets. *Anais do III Dataset Showcase Workshop*. SBC, 89–100.

Chen, Jiawei, Janusz Konrad, and Prakash Ishwar, 2018: Vgan-based image representation learning for privacy-preserving facial expression recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1570–1579.

Choi, Yunjey, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, 2018: Stargan: unified generative adversarial networks for multi-domain image-to-image translation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Cormode, Graham, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang, 2018: Privacy at scale: local differential privacy in practice. *Proceedings of the 2018 International Conference on Management of Data*, 1655–1658.

Differential Privacy Team, Apple, 2017: Learning with privacy at scale differential.

Ding, Bolin, Janardhan Kulkarni, and Sergey Yekhanin, 2017: Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, **30**.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., 2020: An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dwork, Cynthia, 2008: Differential privacy: a survey of results. *International conference on theory and applications of models of computation*. Springer, 1–19.

Dwork, Cynthia, and Aaron Roth, Jan. 2013: The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, **9**. DOI: 10.1561/0400000042.

Edwards, Harrison, and Amos J. Storkey, 2016: Censoring representations with an adversary. *CoRR*, **abs/1511.05897**.

El Emam, Khaled, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin, 2011: A systematic review of re-identification attacks on health data. *PloS one*, **6**, e28071.

Erlingsson, Úlfar, Vasyl Pihur, and Aleksandra Korolova, 2014: Rappor: randomized aggregatable privacy-preserving ordinal response. *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, 1054–1067.

Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart, 2015: Model inversion attacks that exploit confidence information and basic countermeasures. *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1322–1333.

Gambs, Sébastien, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez, 2014: De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, **80**, 1597–1614.

Ganju, Karan, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov, 2018: Property inference attacks on fully connected neural networks using permutation invariant representations. *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 619–633.

Gong, Neil Zhenqiang, and Bin Liu, 2016: You are who you know and how you behave: attribute inference attacks via users' social friends and behaviors. *25th USENIX Security Symposium (USENIX Security 16)*, 979–995.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, 2014: Generative adversarial nets. *Advances in neural information processing systems*, **27**.

Hardt, Moritz, Katrina Ligett, and Frank Mcsherry, 2012: A simple and practical algorithm for differentially private data release. *Advances in Neural Information Processing Systems*, **25**, 2339–2347.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 2016: Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Irvin, Jeremy, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al., 2019: Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI conference on artificial intelligence*. Volume 33. 01, 590–597.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, 2017: Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Kasiviswanathan, Shiva Prasad, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith, 2011: What can we learn privately? *SIAM Journal on Computing*, **40**, 793–826.

Kim, Taeksoo, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim, 2017: Learning to discover cross-domain relations with generative adversarial networks. *International conference on machine learning*. PMLR, 1857–1865.

Kosinski, Michal, David Stillwell, and Thore Graepel, 2013: Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, **110**, 5802–5805.

Lei, Jing, 2014: Classification with confidence. *Biometrika*, **101**, 755–769.

Liu, Ziwei, Ping Luo, Xiaogang Wang, and Xiaoou Tang, 2015: Deep learning face attributes in the wild. *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Liu, Ziwei, Ping Luo, Xiaogang Wang, and Xiaoou Tang, 2018: Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, **15**, 11.

Mangat, Naurang S, 1994: An improved randomized response strategy. *Journal of the Royal Statistical Society: Series B (Methodological)*, **56**, 93–95.

Mao, Xudong, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, 2017: Least squares generative adversarial networks. *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Melis, Luca, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov, 2019: Exploiting unintended feature leakage in collaborative learning. *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 691–706.

Narayanan, Arvind, Elaine Shi, and Benjamin IP Rubinstein, 2011: Link prediction by de-anonymization: how we won the kaggle social network challenge. *The 2011 International Joint Conference on Neural Networks*. IEEE, 1825–1834.

Narayanan, Arvind, and Vitaly Shmatikov, 2006: How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*.

Narayanan, Arvind, and Vitaly Shmatikov, 2008: Robust de-anonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 111–125.

Nasr, Milad, Reza Shokri, and Amir Houmansadr, 2019: Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. *2019 IEEE symposium on security and privacy (SP)*. IEEE, 739–753.

Odena, Augustus, Christopher Olah, and Jonathon Shlens, 2017: Conditional image synthesis with auxiliary classifier gans. *International conference on machine learning*. PMLR, 2642–2651.

Pittaluga, Francesco, Sanjeev Koppal, and Ayan Chakrabarti, 2019: Learning privacy preserving encodings through adversarial training. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 791–799.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., 2021: Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*. PMLR, 8748–8763.

Ren, Zhongzheng, Yong Jae Lee, and Michael S Ryoo, 2018: Learning to anonymize faces for privacy preserving action detection. *Proceedings of the european conference on computer vision (ECCV)*, 620–636.

Rigaki, Maria, and Sebastian Garcia, 2020: A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646*.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox, 2015: U-net: convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., 2015: Imagenet large scale visual recognition challenge. *International journal of computer vision*, **115**, 211–252.

Theil, Henri, 1971: Applied economic forecasting.

Warner, Stanley L, 1965: Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63–69.

Xiao, Taihong, Yi-Hsuan Tsai, Kihyuk Sohn, Manmohan Chandraker, and Ming-Hsuan Yang, 2020: Adversarial learning of privacy-preserving and task-oriented representations. *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 34. 07, 12434–12441.

Xiong, Zuobin, Wei Li, Qilong Han, and Zhipeng Cai, 2019: Privacy-preserving auto-driving: a gan-based approach to protect vehicular camera data. *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 668–677.

Ying, Zuobin, Yun Zhang, and Ximeng Liu, 2020: Privacy-preserving in defending against membership inference attacks. *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, 61–63.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A Efros, 2017: Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

초 록

지역적 차등 정보 보안(Local Differntial Privacy, 이하 LDP)은 널리 알려진 보안에 대한 엄밀한 수학적 정의로, 민감한 데이터에 관해 정량화된 강력한 정보 보안을 보장한다. 하지만 LDP를 이루는 근본적인 메커니즘인 무작위 응답(Randomized Response, 이하 RR)은 테이블 데이터와 같은 구조화된 데이터를 위해 만들어졌으므로 널리 알려진 LDP 알고리즘들은 이미지와 같은 비구조화된 데이터에는 적용하기 어렵다는 단점이 있다. 본 연구에서는 해당 단점을 보완하기 위해 이미지 인코딩 상에서 다른 시각적 특징들은 유지하면서 선택된 민감한 정보들의 보안을 유지하는 LDP 프레임워크를 제안한다. 제안된 프레임워크는 적대적 학습을 통해 생성된 전이 모델을 이용해 RR 메커니즘을 모사함으로써 다른 LDP 알고리즘들에도 쉽게 적용이 가능하다는 장점이 있다. 본 논문에서는 문제 상황을 엄밀히 정의하고 제안된 프레임워크가 민감 정보를 탈취하려는 목적을 가진 잠재적 적대자로부터 정보를 보호할 수 있다는 것을 입증한다. 또한 본 논문에서는 실험적 결과를 통해 제안된 모델이 다른 기존 모델들에 비해 데이터의 잠재적인 미래 작업들에 최대한 영향을 적게 끼치면서 정보를 보호할 수 있다는 것을 보인다.

학 번: 2020-20277