This WACV 2023 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Interpreting Disparate Privacy-Utility Tradeoff in Adversarial Learning via Attribute Correlation

Likun Zhang[†,‡], Yahong Chen[†,‡], Ang Li[§], Binghui Wang[¶], Yiran Chen[§],
Fenghua Li[†,‡], Jin Cao[♮], Ben Niu[†*]
[†]*Institute of Information Engineering, CAS, Beijing, China*
[‡]*School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China*
[§]*Department of Electrical and Computer Engineering, Duke University, Durham, USA*
[¶]*Department of Computer Science, Illinois Institute of Technology, Chicago, USA*
[♮]*School of Cyber Engineering, Xidian University, Xi'an, China*
{zhanglikun, chenyahong, niuben}@iie.ac.cn

Abstract

Adversarial learning is commonly used to extract latent data representations which are expressive to predict the target attribute but indistinguishable in the privacy attribute. However, whether they can achieve an expected privacy-utility tradeoff is of great uncertainty. In this paper, we posit it is the complex interaction between different attributes in the training set that causes disparate tradeoff results. We first formulate the measurement of utility, privacy and their tradeoff in adversarial learning. Then we propose the metrics of Statistical Reliability (SR) and Feature Reliability (FR) to quantify the relationship between attributes. Specifically, SR reflects the co-occurrence sampling bias of the joint distribution between two attributes. Beyond the explicit dependence, FR exploits the intrinsic interaction one attribute exerts on the other via exploring the representation disentanglement. We validate the metrics on CelebA and LFW dataset with a suite of target-privacy attribute pairs. Experimental results demonstrate the strong correlations between the metrics and utility, privacy and their tradeoff. We further conclude how to use SR and FR as a guide to the setting of the privacy-utility tradeoff parameter.

1. Introduction

Deep learning systems have been widely deployed in many high-stakes applications, such as face recognition and commercial analysis, which are heavily dependent on the training data collected from users to achieve satisfying performance. To mitigate users' privacy concern, a common solution is extracting the feature representation of raw data, which should satisfy two goals: 1) indistinguishable in the privacy attribute; 2) expressive to predict the target attribute. For example, in a gender classification system, the image representation should be informative for the model to recognize whether a person wears eyeglasses while the sensitive attributes such as ethnicity or gender should be hidden. These two goals are simultaneously achieved via adversarial learning, by adding a privacy regularization term to the target loss objective when training the feature extractor.

Unfortunately, the quality of representation is often observed to suffer from great uncertainty. When implementing the adversarial learning, the privacy and utility objectives are simply weighted, which is difficult for users to find an optimal tradeoff parameter. Even with a same tradeoff parameter, the tradeoff performance varies with different target-privacy attribute pairs. Specifically, the classification accuracy of the target attribute sometimes suffers from a significant drop and the privacy attribute is still leaked, while sometimes the privacy goal is easily realized without too much utility loss.

In this paper we posit an assumption to interpret such phenomenon: the disparate privacy-utility tradeoff comes from the complex correlation between the target and privacy attribute. Theoretically, gradients of different goals may interfere and multiple summed losses may make the optimization landscape more difficult. When task objectives do not interfere much with each other, it may lead to an ideal tradeoff that the utility of target task suffers no or just a slight drop while the classification accuracy of the privacy task is close to a random guess. In order to quantify such relationship, existing works rely on the effect of one task's gradients on another task's loss of multi-task learning and the model transferability of transfer learning [5, 19], but they fail to

*Corresponding author.

reveal the fundamental reasons why the disparate tradeoff differs in tasks.

Different from the result-oriented relationships, we explore the dominant factors that determine the tradeoff variations, independent from the learning results. In detail, we propose two correlation metrics: The first one is **Statistical Reliability (SR)** which reflects the co-occurrence sampling bias of the training set. Specifically, it refers to the case where an attribute value is always found to co-occur with another attribute value, thus the demographic groups distinguished by a attribute have a large overlap with the groups distinguished by another attribute. It would be especially problematic for datasets involving sensitive attributes related to race, gender, religion, etc., since it could be mistakenly used to support a reckless or even discriminative prediction. The second one is **Feature Reliability (FR)**, which explores the intrinsic causal correlation between two attributes. To interpret such correlation, we extract a data sample into a disentangled feature representation, where each dimension of the feature representation is an independent variable. Various attributes can be encoded with different attention weights. FR measures the distribution similarity of weights between any two attributes, which reflects how obfuscating one attribute affects the recognition of the other attribute.

Our contributions can be summarized as follows:

- We propose SR and FR metrics to measure the explicit statistical reliability and intrinsic feature reliability between two attributes.They explain the reasons of disparate tradeoff that hiding some attributes may lead to dramatic accuracy drop in classifying a target attribute while hiding another attribute hardly has any impact when applying adversarial learning to generate privacy-preserving data representations for a target classification task.

- We conduct comprehensive experiments with a large number of target-privacy attribute pairs in an adversarial learning framework to validate the impact of SR and FR on the privacy-utility tradeoff. The results demonstrate strong correlations between them, and we further conclude a set of general rules to guide the selection of the optimal privacy-utility tradeoff parameter.

- As SR supports a comprehensive statistical sampling bias analysis of the training set, it can detect the potential unfairness existed in AI systems. FR exploits the intrinsic and disentangled relationship between attributes, it can be used as task similarity to generate task taxonomy to reuse supervision among redundant tasks or solve multiple tasks in one system without piling up the complexity.

The rest of this paper is organized as follows. In Sec. 2, we review related works. Sec. 3 formulates the problem and defines measures of privacy-utility tradeoff. In Sec. 4, we propose SR, FR and their application values in detail. Experimental evaluation results are given in Sec. 5. Limitations and future work are discussed in Sec. 6. Finally, we conclude this paper in Sec. 7.

2. Related Works

2.1. Adversarial Learning

Song *et al.* [18] demonstrated that latent data representations overlearned by deep models reveal sensitive attributes that are not part of the training objectives. To generate privacy-preserving data representations, a typical approach is based on adversarial learning [8, 20, 21]. Edwards and Storkey [2] formulated it as a mini-max game between a privacy discriminator and a target model who try to get over against each other with opposed objectives. Similar methods have been extended to various scenarios, such as census and health records [22], texts [3, 11], images [4, 6, 17] and sensor data of wearables [7]. The privacy objectives of adversarial learning could also be formulated as minimizing the mutual information between the representation and the sensitive attribute[16]. Li *et al.* [9] proposed a task-independent adversarial learning framework by maximizing the mutual information between the representations and the raw data while minimizing the mutual information between the representation and the privacy attribute.

2.2. Task Correlation

In terms of different attribute classification tasks, a straightforward way to measure their correlation is via the statistical features of the joint distribution between attributes. Melis *et al.* [15] calculated Pearson correlation coefficient between labels of the main attribute and the privacy attribute, but Pearson correlation applies to numerical variables. Zhang *et al.* [24] used Cramer's V [1] for categorical-categorical variables, but there is no unified standard for the value of Cramer's V correlation coefficient which should be evaluated via hypothesis test, thus it cannot be directly utilized to compare the correlations of different attribute pairs.

Beyond attribute classification tasks, in the context of transfer learning, Zamir *et al.* [23] computed task affinity matrix based on whether the representation for one task can be sufficiently transferred to train for another task. Fifty *et al.* [5] improved the calculation efficiency of task affinity in a single run by training all tasks together and quantifying the effect of one task's gradients on another task's loss. Since the transfer relationships are not highly predictive of multi-task relationships, Standley *et al.* [19] empirically studied the average change in the performance of two tasks when they are trained as a pair compared to when they are trained separately.

However, the above relationships are result-oriented, calculated based on the actual learning performance. Also, they fail to interpret the fundamental causes why the joint learning results differ for different combination of tasks. Different from the existing works, we focus on the the scenarios of adversarial learning on multi-attribute data, and explore the intrinsic correlation between different attribute classification tasks, independent from the learning process.

3. Preliminaries

3.1. Problem Statement

Given an input x, its target attribute value is $y \in \{0, 1, ..., m-1\}$, and the privacy attribute value is $p \in \{0, 1, ..., n-1\}$. The ultimate goal is to learn a feature extractor E_θ parameterized by θ to encode x into a representation $z = E_\theta(x)$, which satisfies two goals:

- Privacy: indistinguishable in the privacy attribute p;
- Utility: expressive to predict the target attribute y.

For example, in a mask detection model, the latent representations of people's facial image are used to recognize whether they wear a mask, but it should not reveal any deterministic information about their gender privacy. It ends up with a tradeoff between privacy and utility, which is controlled by a parameter λ to adjust the importance weights of these two goals. Accordingly, we define the measurement of **privacy**, **utility** and their tradeoff in Sec. 3.3.

However, the quality of representations suffer from great uncertainty. That means the classification performance of the target attribute sometimes suffers from a significant drop but the privacy attribute is still leaked while sometimes the utility and privacy goals can both be realized. This may be because the different tasks are learned at different rates. Or because one task may dominate the learning leading to poor performance on the other. Furthermore, gradients of different goals may interfere and multiple summed losses may make the optimization landscape more difficult. On the other hand, when task objectives do not interfere much with each other, it may lead to an ideal tradeoff that the utility of target task is maintained or just suffers a slight drop while the classification accuracy of the privacy task is close to a random guess. Intuitively, the disparate tradeoff results heavily depend on the relationship between classification attributes, such as the hair color and skin color, shopping preference and gender.

Assumptions. Thus we propose our assumption that it is the potential correlations between different classification attributes that exert dominant impact on the performance of utility, privacy and their tradeoff. The more similarity two attributes have, the harder to remove the privacy attribute while retaining the target attribute. In Sec. 4, we study these relationships in depth.

3.2. Formulation of Adversarial Learning

In this subsection, we formally describe a typical privacy-preserving adversarial learning framework. To achieve the privacy goal, there is an adversary model A_ψ parametrized by ψ which aims to perfectly predict p from $A_\psi(z)$. Therefore, it minimizes the following loss function:

$$\mathcal{L}_A = \mathcal{J}(A_\psi(E_\theta(x)), p), \tag{1}$$

where $\mathcal{J}(\cdot, \cdot)$ denotes the cross-entropy loss function. By contrast, the defender aims to fail A_ψ, i.e., maximize $\mathcal{L}_A$. However, this will push the generated feature representation towards the opposite side of the privacy attribute, e.g., $p = 1$ flips to $p = 0$. Therefore, we make the prediction of A_ψ a random guess by increasing the entropy of $A_\psi(z)$. Thus the privacy loss of the defender can be formulated as:

$$\mathcal{L}_D^p = -\mathcal{L}_A - \alpha \mathcal{H}(A_\psi(E_\theta(x))), \tag{2}$$

where $\mathcal{H}$ calculates the entropy and $\alpha > 0$ controls the entropy term. At the same time, to achieve the utility goal, the defender needs to make the target classifier C_ϕ parametrized by ϕ precisely infer y from z. The utility loss can be expressed as:

$$\mathcal{L}_D^u = \mathcal{J}(C_\phi(E_\theta(x)), y). \tag{3}$$

Combining the utility and privacy goals, the total objective of the defender can be formulated as:

$$\mathcal{L}_D = \lambda \mathcal{L}_D^p + (1 - \lambda)\mathcal{L}_D^u, \tag{4}$$

where $\lambda \in [0, 1]$ is the tradeoff parameter. A larger λ indicates a stronger privacy guarantee, while a smaller λ allows more utility retained in the extracted features. During the optimization, A_ψ and E_θ, C_ϕ are updated alternatively to minimize $\mathcal{L}_A$ and $\mathcal{L}_D$, respectively.

3.3. Measures of Privacy and Utility

We describe how to fairly measure the privacy and utility in adversarial learning in detail. Specifically, we randomly split the total dataset D into a training set D_0 and a testing set D_1. Given a privacy attribute p, we use D_0 to train a privacy-preserving feature extractor E utilizing the adversarial learning framework described in Sec. 3.2. Then we randomly select 50% samples from D_0 and take them as the augmentation dataset D_{aug} to train a validation classifier (VC) for each target attribute and privacy attribute using feature representations $E_{x \in D_{aug}}(x)$. We evaluate each validation classifier on D_1 and denote its accuracy performance as $f_a(\cdot)$.

To fairly evaluate the utility gain and privacy loss, we also use D_0 to normally train a baseline classification model (BC) without any privacy-preserving objective for each attribute as the best classification performance. The accuracy

performance of the normal classification model for each attribute is tested with D_1 as well, denoted as $f_n(\cdot)$. In addition, the accuracy of each attribute when making random guess is denoted as $f_r(\cdot)$.

So both the closeness to a random guess and the deviation from the normal training performance should be considered when evaluating the privacy and utility. Thus we measure the privacy leakage level for attribute p as follows:

$$M_p = \frac{f_a(p) - f_r(p)}{f_n(p) - f_r(p)} \tag{5}$$

where $f_*(i)$ refers to accuracy performance of the corresponding classifier for the attribute i. The **lower** M_p, the closer to a random guess and better privacy guarantee.

Similarly, the utility performance on y is measured as:

$$M_u = \frac{f_a(y) - f_r(y)}{f_n(y) - f_r(y)} \tag{6}$$

The **larger** M_u, the less utility drop.

Accordingly, the privacy-utility tradeoff is calculated as:

$$T = \frac{M_p}{\delta + M_u} \tag{7}$$

where $\delta = 0.0001$, in case of the zero-division error. The **lower** T, the better tradeoff result, i.e., removing the privacy attribute has less negative impact on the target attribute.

4. Methodology

We analyze the possible causes of the disparate impact on privacy-utility tradeoff among diverse adversarial learning tasks. Only when we uncover the decisive influencing factors, could we design effective metrics to reflect the relationship between privacy and target attributes.

4.1. Statistical Reliability

Intuitively, a direct cause is the statistical dependence between the privacy attribute and target attribute in the training set caused by imbalanced and insufficient sampling. For two different attributes a and b, we denote their classifiers as M_1 and M_2 which use a joint training dataset. If a and b are statistically related in the training dataset, a certain label of the attribute a will be always found to co-occur with a certain value of b such as "makeup" and "female". As a result, the consistency between the separation hyper-planes of M_1 and M_2 are correlated with the statistical reliability between y and p.

We first define the statistical reliability between two binary attributes $a \in \{0, 1\}$ and $b \in \{0, 1\}$, where $a = 1$ means a sample has the attribute a otherwise the sample does not. The **Statistical Reliability (SR)** between the attribute a and b is defined as:

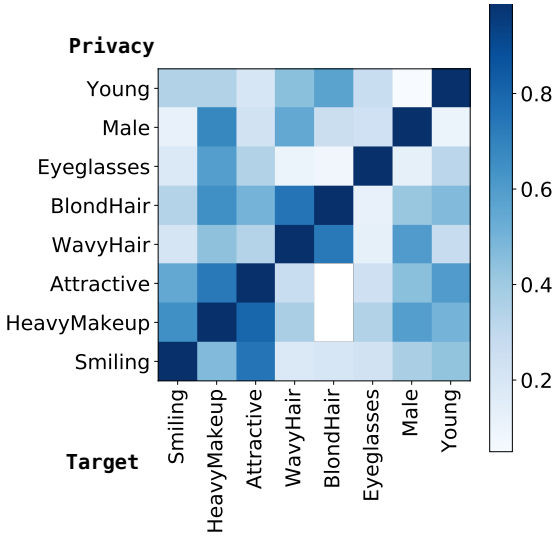
$$SR(a, b) = 1 - \frac{4(C_{00} + C_{11})(C_{10} + C_{01})}{N^2}, \tag{8}$$

where C_{ij} is the number of data samples labeled with $a = i, b = j$, and N refers to the total size of data samples. When $C_{00} + C_{11} = C_{10} + C_{01}, SR(a, b) = 0$, there is no statistic correlation between a and b.

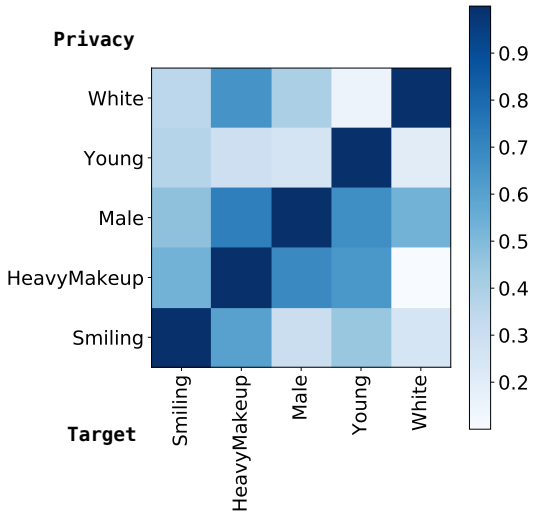
Then we extend the metric to attributes with multiple labels. If an attribute a has $n > 2$ labels, it could be extended to n binary attributes $\{a_0, a_1, ..., a_{n-1}\}$. If a sample has the attribute i, we define $a_i = 1$, otherwise $a_i = 0$. Without generality, **SR** between the attributes a with n labels and b with m labels can be measured as follows:

$$SR(a, b) = \max_{\substack{i=0,1,...,n-1 \\ j=0,1,...,m-1}} \{SR(a_i, b_j)\}. \tag{9}$$

Obviously, $SR(a, b) \in [0, 1]$. The larger SR, the more significant statistical reliability.



(a) FR Correlation of CelebA



(b) FR Correlation of LFW

Figure 1. FR Correlation Value Matrix(the horizontal/vertical axis refers to the target/privacy attribute)

4.2. Feature Reliability

The disparate tradeoff caused by statistical reliability can be mitigated by re-sampling or re-weighting the training data, however, there is also intrinsic feature reliability among different attributes in the level of latent features. To interpret such reliability, assume each dimension of the feature representation is a variable independent from the others. The classifier thus is optimized to encode those variables which exert causal effect with larger weights to make the final prediction. Some attributes can be encoded by common variables which means they share common causes. For example, hair color could still be inferred from skin color because they jointly share some endogenous features. Thus it is easy to understand why removing the information of one attribute may result in negative impact on the recognition of another attribute if they are closely related. In contrast, if two attributes have lower intrinsic correlation but explicit statistical bias, removing one attribute has little effect on the other.

To estimate such intrinsic correlation between attributes, we borrow the notion of disentangled feature representation [25] where each dimension is independent from the others.

Notations. $\mathbb{R}^{m_X}$ denotes the space of raw data, $\mathbb{R}^{m_Y}$ denotes the output space and $\mathbb{R}^{m_Z}$ denotes the feature representation space. m_X, m_Y, m_Z correspond to the dimensions of space X, Y, Z, respectively. $f : X \to Z$ denotes the representation function, and $g : Z \to Y$ is the prediction function parameterized by W. We have N raw data samples $X \subset \mathbb{R}^{N \times m_X}$ with labels $Y \subset \mathbb{R}^{N \times m_Y}$. The representations learned by f are donated as $Z \subset \mathbb{R}^{N \times m_Z}$. The i-th variable in the representation space is donated as $Z_{:,i}$.

First, we train an auto-encoder $f \circ h$ using the method of StableNet [25], such that all input samples X can be encoded with disentangled feature representations $Z = f(X)$, where any pair of variables $Z_{:,i}$ and $Z_{:,j}$ are independent of each other. The input sample can be reconstructed through the decoder by minimizing $\|X - h(Z)\|_2$, ensuring Z embeds key information of X. Then, each classification attribute Y^k can be inferred from a classifier $Y^k = g^k(Z)$ parameterized by W^k. Variables in $\mathbb{R}^{m_Z}$ can be regarded as a collection of informative atomic-level features. W^k indicates how much attention a target task Y^k should pay to these variables when making a decision for the output. For each attribute pair a and b with m and n possible values, the shape of W^a and W^b could be the same when their classifiers g^a and g^b are of the same structure whose input length is $|m^z|$ and the output length $m_Y = \max\{m, n\}$. If we use a fully connected layer as the classifier, the shape of W^a and W^b are identical to $m^z \times m_Y$.Thus we can easily calculate the **Feature Reliability (FR)** according to the distribution similarity of W between any two attributes. Given a target attribute a, the extent of dependency a relies on the other attribute b can be defined as follows:

$$FR(a, b) = \frac{W^a \cdot W^b}{\|W^a\|_2^2} \tag{10}$$

It is noteworthy that FR describes how the recognition of the attribute a relies on the attribute b, so FR value is not necessarily symmetric, i.e., $FR(a, b) \neq FR(b, a)$.

4.3. Using SR and FR

We envision the proposed SR and FR metrics to be used for the following purposes:

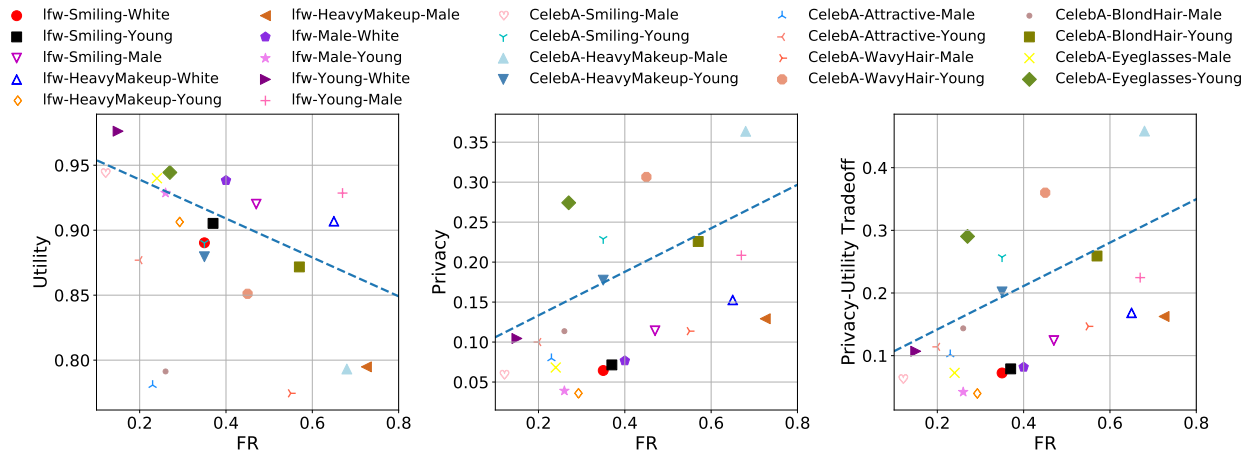
- The proposed correlation metrics could be used to guide the selection of the privacy-utility tradeoff parameter in adversarial learning to reach the expected tradeoff. Specifically, the optimal parameter differs as SR and FR vary for different attribute pairs. Detailed selection rules will be presented in Sec. 5.

- As the proposed SR supports a comprehensive statistical bias analysis of the training set, it can be a prompt of potential unfairness since biased dataset is a main cause of unfairness in AI systems [12]. In practice, it has been demonstrated that many public datasets such as COCO [13], show a notable sampling bias that most images in shopping or cooking scenarios are connected with females while those coach images in sport scenarios are mostly males, which is a reflection of unwanted social bias. It would be especially problematic for datasets involving sensitive attributes related to race, gender, religion, age, status, physical traits, etc., because the statistical tendency could be mistakenly used to support a reckless and discriminative decision during model training.

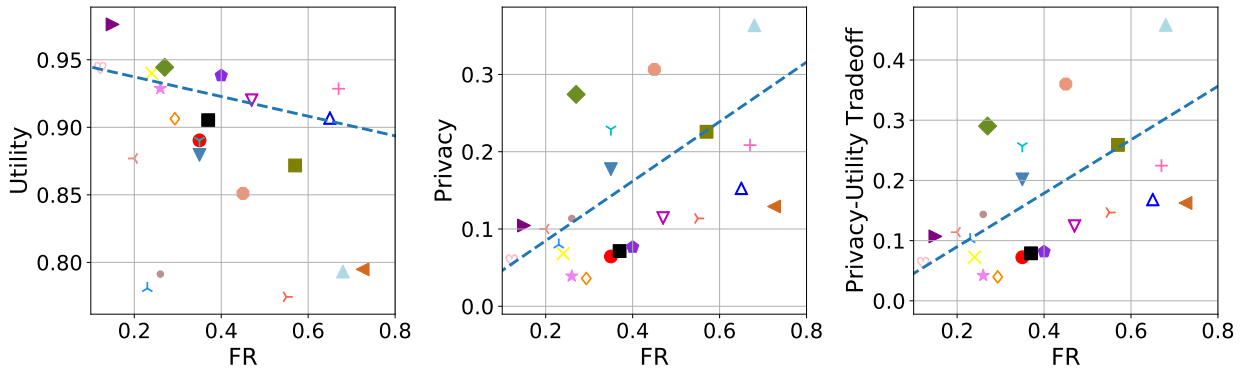
- A model aware of task relationships demands less supervision, uses less computation and behaves in more predictable ways [5, 19]. Since the proposed FR exploits the essential relationship between tasks, it can be used as task similarity, which creates an association and/or causation relationship among tasks with high correlation. It implies that highly correlated tasks mutually reinforce each other when putting them together in multi-task learning. Thus it could help generate task taxonomy to reuse supervision among redundant tasks or solve multiple tasks in one system without piling up the complexity.

5. Empirical Evaluation

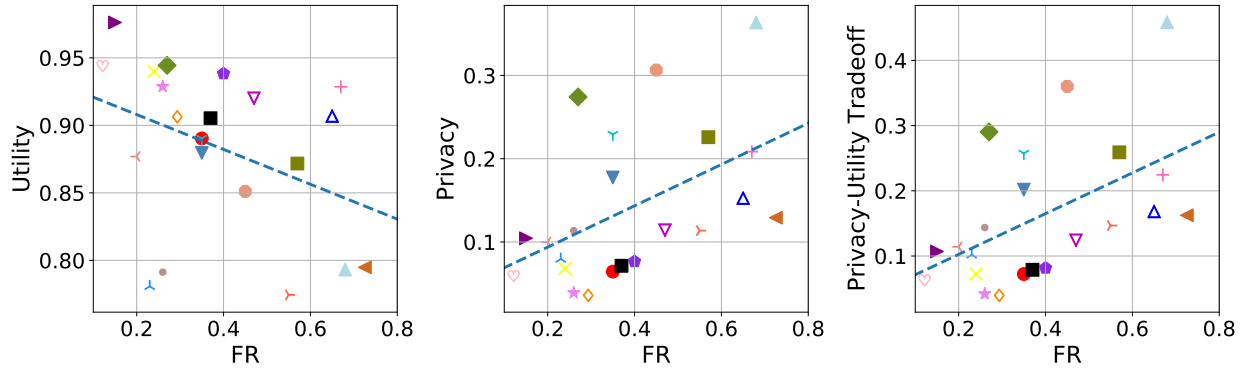
In this section, we implement the adversarial learning framework stated in Sec. 3.2 on PyTorch v1.4.0 with a Nvidia Tesla T4 GPU, following the evaluation method proposed by Li *et al.* [10]. Our goal is to answer the following questions through comprehensive experimental analysis:



(a) $\lambda = 0.3$



(b) $\lambda = 0.5$



(c) $\lambda = 0.7$

Figure 2. Impact of FR Correlation on Utility and Privacy for LFW and CelebA Dataset

First, when removing statistical reliability, how does FR affect the privacy-utility tradeoff results?

Second, when training model using the dataset with different extent of statistical reliability, how does SR affect the privacy-utility tradeoff results? Will the impact of SR subject to FR?

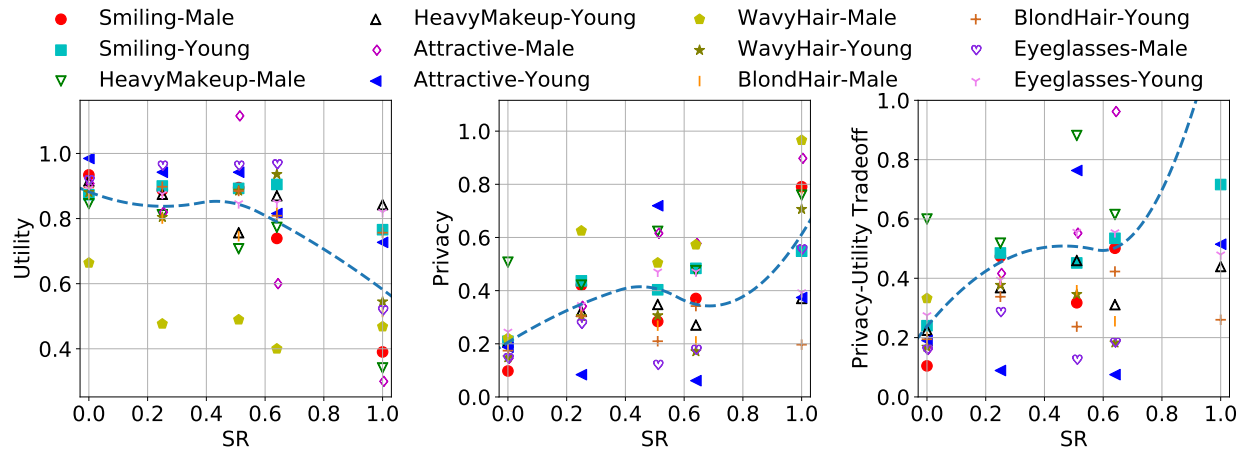
Third, how do we use SR and FR metrics to instruct the setting of λ to meet the expected utility and privacy requirement?

5.1. Experimental Setup

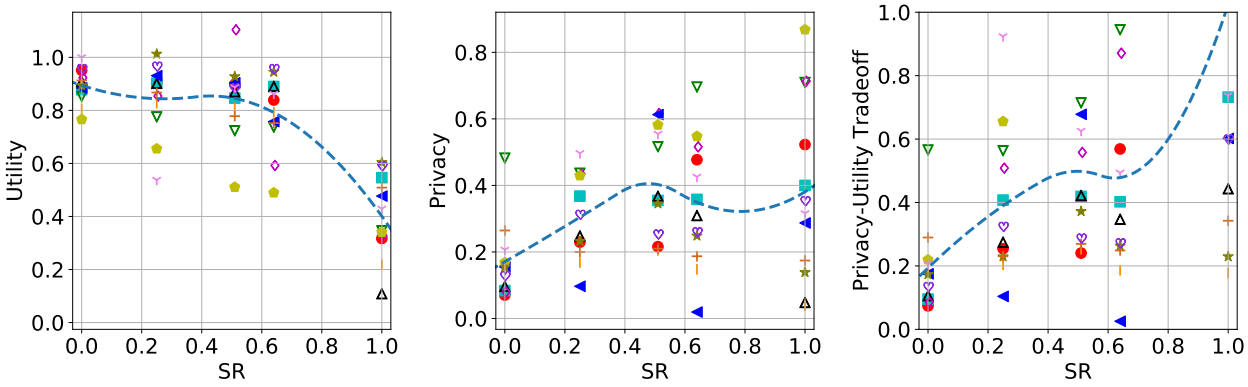
Datasets. We adopt two multi-attribute datasets:

- CelebA dataset[1] contains over $200k$ facial images labeled with 40 diverse attributes. We select "Smiling, Wavy hair, Blonde hair, Heavy makeup, Eyeglasses, Attractive" as the target attribute, and "Male, Young" as the privacy attribute.

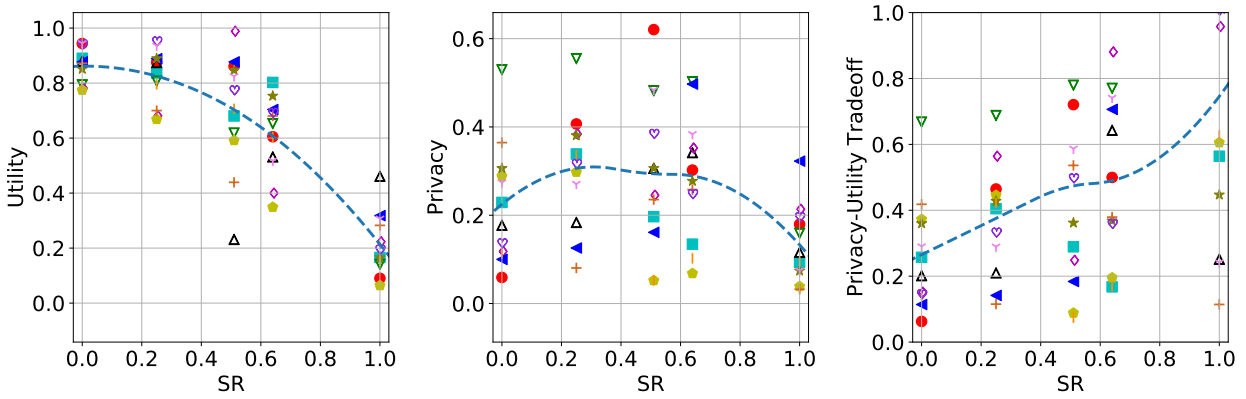
[1] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html



(a) $\lambda = 0.3$



(b) $\lambda = 0.5$



(c) $\lambda = 0.7$

Figure 3. Impact of SR Correlation on Utility and Privacy for CelebA Dataset

- LFW dataset[2] contains $13,244$ facial images with 73 attributes. We take "Smiling, Heavy makeup, Male" as the target attribute, "White, Male, Young" as the privacy attribute.

Implementation details. For fairness, we remove the statistical biases of the testing set by setting $(C_{00} : C_{10} : C_{01} : C_{11}) = (1 : 1 : 1 : 1)$. For more details, please refer to the Appendix.

[2]http://vis-www.cs.umass.edu/lfw/, attribute annotations are in [14]

5.2. Validation Results of FR

To answer the first question, we first set the value of SR to be 0 by re-sampling the training set as $(C_{00} : C_{10} : C_{01} : C_{11}) = (1 : 1 : 1 : 1)$ to study the impact of FR correlation independently. The estimated FR correlations between all permutation pairs of selected attributes are shown as a two-dimension matrix in Fig. 1.The effect of y on p may not be

necessarily identical to that of p on y, thus the matrix is not symmetric along the diagonal.

It is noteworthy that the attribute similarity is to some extent consistent with semantic similarity in terms of human perception. For example, the "HeavyMakeup-Male" owns the highest semantic correlation among all the chosen attribute pairs in our common sense, since the behavior of making up is more likely associated with females, and the estimated FR correlation is also as high as 0.7 in LFW dataset and 0.67 in CelebA. However, there is also exception. "Smiling-HeavyMakeup" seems to have no obvious association, but its FR value exceeds 0.6, which reflects the difference in perception of humans and neural networks.

The effects of FR on utility, privacy and their tradeoff are presented in Fig. 2. The blue dashed line in each figure is the overall trend line. In each sub-figure, the horizontal coordinate axis is the FR correlation calculated in Fig. 1. The vertical coordinate axis of "Utility, Privacy, and Privacy-Utility Traddoff" refers to M_u, M_p, T metrics, respectively. Specifically, for a highly correlated pair "HeavyMakeup-Male" with $\lambda = 0.5$, the target classification accuracy is about 10% lower than the baseline, while its privacy leakage M_p is a lot higher than that of attribute pairs with lower FR correlations such as "Smiling-Male". When $\lambda = 0.7$, for attribute pairs with FR smaller than 0.4, the classification accuracy of the target attribute drops largely while the privacy preservation effectiveness has no significant improvement. On average, the Pearson correlation coefficient between FR and M_u is about -0.63, 0.58 between FR and M_p, 0.68 between FR and T.

5.3. Validation Results of SR

To answer the second question, we impose SR to be $0, 0.25, 0.51, 0.64, 1$ by weighted re-sampling of the training data with $(C_{00} : C_{10} : C_{01} : C_{11}) = (1 : 1 : 1 : 1), (1 : 3 : 3 : 1), (1 : 6 : 6 : 1), (1 : 9 : 9 : 1), (0 : 1 : 1 : 0)$, respectively. Fig.3 shows the impact of SR correlation on utility gain M_u, privacy loss M_p and their tradeoff T. When SR≤ 0.3, the utility and privacy-utility tradeoff have no significant change under all λ, but λ has a positive correlation with the privacy preservation effectiveness. When SR gets larger than 0.4, it appears to be dramatic falls in utility gain and privacy preservation effectiveness. When SR further increases larger than 0.6, it has a sharp decline in utility regardless of λ, however, the privacy leakage begins to decrease when $\lambda = 0.7$.

It is noteworthy that the changing trend of some attribute pairs with high FR correlation such as "Attractive-Male" and "HeavyMakeup-Male" is inconsistent with the overall trend line, i.e., larger SR takes no significant negative impact on the utility performance and privacy leakage compared with SR$= 0$. A possible explanation is that these attribute pairs are naturally co-founded with high SR values, thus the disparate tradeoff results are not caused by biased sampling, but their intrinsic feature reliability.

Guide to the setting of λ: When FR≤ 0.5, SR takes a dominant effect to adjust the value of λ. When SR≥ 0.6 in the training dataset, setting a larger tradeoff parameter such as $\lambda = 0.7$ is recommended to satisfy the privacy preservation goal. When SR ranges from 0.4 to 0.6, $\lambda \leq 0.5$ is recommended to achieve a better privacy-utility tradeoff with no significant utility drop but modest privacy leakage compared to $\lambda = 0.7$. When SR is smaller than 0.4, users can set λ according to their privacy preservation requirement. If they put greater emphasis on privacy, a larger value such as $\lambda > 0.5$ is recommended.

When FR≥ 0.5, FR has more obvious influence than SR. It is recommended to set the value of λ larger than 0.5 to meet the demand of privacy preservation. If the user pays more attention to the utility and agrees to sacrifice the privacy to some extent, λ should be set lower than 0.5 to improve the utility as much as possible.

6. Discussion

The study of attribute correlation casts light on the interpretability of adversarial learning, but our work still has some limitations. First, we only consider one privacy goal, in the future, we will use SR and FR to generate taxonomy for multiple privacy attributes, and set priority weights for them accordingly. Second, we focus on attribute classification tasks. Third, the calculation of FR is based on the existing disentangled representation learning, which is still an ongoing research. We will keep exploring model disentanglement to optimize the estimation of FR.

7. Conclusions

In this paper, we derive the correlation between attributes to interpret disparate privacy-utility tradeoff when applying adversarial learning to extract latent representations which are expressive for the target attribute but indistinguishable in the privacy attribute. To quantify the correlation, we propose Statistical Reliability(SR) and Feature Reliability(FR). SR measures the co-occurrence distribution bias of two attributes, while FR estimates the intrinsic causal effect one exerts on the other. We implement the adversarial learning on various target-privacy attribute pairs to validate the effectiveness of SR and FR. Results demonstrate they are highly predictive of the privacy-utility tradeoff results. Accordingly, we further conclude general rules to instruct the setting of the privacy-utility tradeoff parameter.

8. Acknowledgements

This work is supported by the National Key R&D Program of China (2021YFB3100300), and the National Natural Science Foundation of China (61932015).

References

[1] Harald Cramer. *Mathematical Methods of Statistics*. Princeton University Press, 1999.

[2] Harrison Edwards and Amos J. Storkey. Censoring representations with an adversary. In *Proc. of ICLR*, 2016.

[3] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proc. of ACL EMNLP*, 2018.

[4] Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel. Learning anonymized representations with adversarial neural networks. *CoRR*, abs/1802.09386, 2018.

[5] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. In *Proc. of NIPS*, 2021.

[6] Jihun Hamm. Minimax filter: Learning to preserve privacy from inference attacks. *J. Mach. Learn. Res.*, 18:129:1–129:31, 2017.

[7] Yusuke Iwasawa, Kotaro Nakayama, Ikuko Yairi, and Yutaka Matsuo. Privacy issues regarding the application of dnns to activity-recognition using wearables and its countermeasures by use of adversarial training. In *Proc. of IJCAI*, 2017.

[8] Jinyuan Jia and Neil Zhenqiang Gong. Attriguard: A practical defense against attribute inference attacks via adversarial machine learning. In *Proc. of USENIX Security*, 2018.

[9] Ang Li, Yixiao Duan, Huanrui Yang, Yiran Chen, and Jianlei Yang. TIPRDC: task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations. In *Proc. of ACM SIGKDD*, 2020.

[10] Fenghua Li, Hui Li, Ben Niu, and Jinjun Chen. Privacy computing: Concept, computing framework, and future development trends. *Engineering*, 5(6):1179–1192, 2019.

[11] Yitong Li, Timothy Baldwin, and Trevor Cohn. Towards robust and privacy-preserving text representations. In *Proc. of ACL ACL*, 2018.

[12] Yi Li and Nuno Vasconcelos. REPAIR: removing representation bias by dataset resampling. In *Proc. of IEEE CVPR*, 2019.

[13] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. of Springer ECCV*, 2014.

[14] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. of IEEE ICCV*, 2015.

[15] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *Proc. of IEEE S&P*, 2019.

[16] Seyed Ali Osia, Ali Taheri, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Hamid R. Rabiee. Deep private-feature extraction. *IEEE Trans. Knowl. Data Eng.*, 32(1):54–66, 2020.

[17] Francesco Pittaluga, Sanjeev J. Koppal, and Ayan Chakrabarti. Learning privacy preserving encodings through adversarial training. In *Proc. of IEEE WACV*, 2019.

[18] Congzheng Song and Vitaly Shmatikov. Overlearning reveals sensitive attributes. In *Proc. of ICLR*, 2020.

[19] Trevor Standley, Amir Roshan Zamir, Dawn Chen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *Proc. of PMLR ICML*, 2020.

[20] Zhibo Wang, Xiaowei Dong, Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, and Kui Ren. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proc. of IEEE CVPR*, 2022.

[21] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proc. of Springer ECCV*, 2018.

[22] Qizhe Xie, Zihang Dai, Yulun Du, Eduard H. Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Proc. of NIPS*, 2017.

[23] Amir Roshan Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proc. of IJCAI*, 2019.

[24] Wanrong Zhang, Shruti Tople, and Olga Ohrimenko. Leakage of dataset properties in multi-party machine learning. In *Proc. of USENIX Security*, 2021.

[25] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *Proc. of IEEE CVPR*, 2021.