Adversarial Learning of Privacy-Preserving and Task-Oriented Representations

Taihong Xiao\(^1\), Yi-Hsuan Tsai\(^2\), Kihyuk Sohn\(^2\)\(^*\), Manmohan Chandraker\(^2\)\(^{,3}\), Ming-Hsuan Yang\(^1\)

\(^1\)University of California, Merced
\(^2\)NEC Laboratories America
\(^3\)University of California, San Diego

Abstract

Data privacy has emerged as an important issue as data-driven deep learning has been an essential component of modern machine learning systems. For instance, there could be a potential privacy risk of machine learning systems via the model inversion attack, whose goal is to reconstruct the input data from the latent representation of deep networks. Our work aims at learning a privacy-preserving and task-oriented representation to defend against such model inversion attacks. Specifically, we propose an adversarial reconstruction learning framework that prevents the latent representations decoded into original input data. By simulating the expected behavior of adversary, our framework is realized by minimizing the negative pixel reconstruction loss or the negative feature reconstruction (i.e., perceptual distance) loss. We validate the proposed method on face attribute prediction, showing that our method allows protecting visual privacy with a small decrease in utility performance. In addition, we show the utility-privacy trade-off with different choices of hyperparameter for negative perceptual distance loss at training, allowing service providers to determine the right level of privacy-protection with a certain utility performance. Moreover, we provide an extensive study with different selections of features, tasks, and the data to further analyze their influence on privacy protection.

Introduction

As machine learning (ML) algorithms powered by deep neural networks and large data have demonstrated an impressive performance in many areas across natural language (Bahdanau, Cho, and Bengio 2015; Wu et al. 2016), speech (Oord et al. 2016) and computer vision (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016), there have been increased interests in ML-as-a-service cloud services. These systems demand frequent data transmissions between service providers and their customers to train ML models, or users to evaluate their data. For example, customers who want to develop face recognition system may share the set of images containing people of interest with cloud service providers (CSPs). Facial expression based recommendation system may ask users to upload photos. Unfortunately, these processes are exposed to serious privacy risks. Data containing the confidential information shared from the customers can be misused by the CSPs or acquired by the adversary. The damage is critical if the raw data is shared with no encryption strategy.

An alternative solution to protect privacy is to encode data using deep features. While these approaches are generally more secure than raw data, recent advances in model inversion (MI) techniques (Mahendran and Vedaldi 2015; Dosovitskiy and Brox 2016b; Zhang, Lee, and Lee 2016; Dosovitskiy and Brox 2016a) call the security of deep features into question. For example, (Dosovitskiy and Brox 2016a) shows that adding deep image prior (e.g., perceptual similarity) allows inversion from low-, mid-, as well as high-level features.

In this work, we study how to learn a privacy-preserving and task-oriented deep representation. We focus on the defense against a black-box model inversion attack, where the adversary is allowed to make unlimited inferences\(^1\) of their representations by

\[^*^\] Now at Google.

Copyright \(\odot\) 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

\(^1\)While our assumption is the most favorable scenario to adversary, in practice, the CSPs can limit the number of inferences. MI attack with limited inference is beyond the scope of this work.

---

Figure 1: Representations of conventional deep learning algorithms are vulnerable to adversary’s model inversion attack (Fredrikson, Jha, and Ristenpart 2015), which raise serious issues on data privacy. Our method learns task-oriented features while preventing private information being decoded into an input space by simulating the adversary’s behavior at training phase via negative reconstruction loss.
own data\textsuperscript{2} to recover input from acquired features of private
customer and user data. As an adversary, we consider an MI
attack based on the neural decoder that is trained to recon-
struct the data using data-feature pairs. Perceptual (Dosovit-
skiy and Brox 2016a) and GAN (Goodfellow et al. 2014)
losses are employed to improve the generation quality. Fi-
nally, we present our adversarial data reconstruction learn-
ing that involves an alternate training of encoder and de-
coder. While the decoder, simulating the adversary’s attack,
is trained to reconstruct input from the feature, the encoder
is trained to maximize the reconstruction error to prevent de-
coder from inverting features while minimizing the task loss.

We explain the vulnerability of deep networks against the
MI attacks in the context of facial attribute analysis with ex-
tensive experimental results. We show that it is difficult to
invent the adversarially learned features and thus the pro-
posed method successfully defends against a few strong in-
version attacks. In this work, we perform extensive exper-
iments by inverting from different CNN layers, with differ-
ent data for adversary, with different utility tasks, with
different weight of loss term, to study their influences on
data privacy. Furthermore, we show the effectiveness of our
method against feature-level privacy attacks by demonstrat-
ing the improved invariance on the face identity, even when
the model is trained with no identity supervision.

The contributions of this work are summarized as follows:

- We propose an adversarial data reconstruction learning to
defend against black-box model inversion attacks, along
with a few strong attack methods based on neural decoder.
- We demonstrate the vulnerability of standard deep fea-
tures and the effectiveness of the features learned with our
method in preventing data reconstruction.
- We show the utility-privacy trade-off with different choice
of hyperparameter for negative perceptual distance loss.
- We perform extensive study of the impacts on the privacy
protection with different layers of features, tasks, and data
for decoder training.

**Related Work**

**Data Privacy Protection**

To protect data privacy, numerous approaches have been
proposed based on information theory (Liang et al.
2009), statistics (du Pin Calmon and Fawaz 2012; Smith
2011), and learnability (Kasiviswanathan et al. 2011). Fur-
thermore, syntactic anonymization methods including k-
anonymity (Sweeney 2002), l-diversity (Machanavajjhala
et al. 2006) and t-closeness (Li, Li, and Venkatasubramanian
2007) are developed. Nevertheless, these approaches pro-
tect sensitive attributes in a static database but do not scale
well to high-dimensional image data. On the other hand, the
concept of differential privacy (Dwork and Nissim 2004;
Dwork et al. 2006) has been introduced to provide formal
privacy guarantee. It prevents an adversary from gaining ad-
tional knowledge by including or excluding an individual

\textsuperscript{2}Although we assume the adversary has no direct access to the
private data used for model training, adversary’s own data is as-
sumed to be representative of them.

**Visual Privacy Protection**

Typical privacy-preserving visual recognition methods aim
to transform the image data such that identity cannot be vi-
ually determined, based on image operations such as Gaussi-
ан blur (Oh et al. 2016), mean shift filtering (Winkler,
Erdélyi, and Rinner 2014), down-scaling, identity obfus-
cation (Oh et al. 2016), and adversarial image perturba-
tion (Oh, Fritz, and Schiele 2017). Although effective in
protecting privacy, these methods have negative impact on
utility. To overcome this limitation, numerous algorithms
have been proposed to complete the utility task based on
transformed data. (Wang et al. 2016) design a method to
improve low-resolution recognition performance via feature
enhancement and domain adaptation. In (Ryoo et al. 2017),
it is demonstrated that reliable action recognition may be
achieved at low resolutions by learning appropriate down-
sampling transformations. Furthermore, trade-offs between
resolution and action recognition accuracy are discussed in
(Dai et al. 2015). Furthermore, (Oh, Fritz, and Schiele 2017)
propose an adversarial method to learn the image perturba-
tion so as to fool the face identity classifier, but the adver-
sarial perturbed images are visually exposing the identity pri-
vacy. Different from these methods, our method learns im-
age features so as to protect the privacy, which could also
maintain the utility performance to some extent.

**Feature and Model Inversion**

There has been a growing interest in stealing the function-
ality from the black-box classifier (Orekondy, Schiele, and
Fritz 2019) or inverting CNNs to better understand what is
learned in neural networks. (Mahendran and Vedaldi 2015)
introduce optimization-based inversion technique that al-
lows low- and mid-level CNN features to be inverted. (Doso-
vitskiy and Brox 2016b; Zhang, Lee, and Lee 2016) suggest
to invert via up-convolution neural network and demonstrate
improved network inversion from mid-level representations.
Nevertheless, this method is less effective in inverting high-
level features. More importantly, recent studies (Dosovit-
skiy and Brox 2016a) show that adding perceptual image
prior allows inversion from high-level representations pos-
sible, exposing a potential privacy issue by inverting deep
representations. However, our method prevents image rep-
resentations from being reconstructed to original images via
adversarial learning.

**Trade-off between Privacy and Utility**

Recently, several methods protect the utility performance
when the information required for the utility task is unrelated
to the private information. For example, (Wu et al. 2018;
Ren, Jae Lee, and Ryoo 2018) disguise original face im-
ages (privacy) without sacrificing the performance of action
recognition (utility), since the information being useful to
action recognition is independent of the face information.
However, there still remains a question of how to better pro-
tect the privacy when the information for the utility task is
Figure 2: An overview of our privacy-preserving representation learning method with adversarial data reconstruction. As in (a), decoder (Dec) is trained to reconstruct an input $X$ from latent encoding $Z = \text{Enc}(X)$ on public data $X_2$. Then, as in (b), we update Enc on private data $X_1$ to generate $Z$ that fools Dec, i.e., prevent reconstructing an input $X$ by Dec, while achieving utility prediction performance via the classifier $f$.

Differences between Our method and Existing methods.

In contrast to prior work, our approach can 1) scale up to high-dimensional image data compared with those methods based on $k$-anonymity, $t$-diversity and $t$-closeness that are often applied to field-structured data (e.g., tabular data); 2) learn image representations as the alternative of raw images while some approaches focus on removing private information in raw images; 3) require no definition of private attribute or entailment performance of additional annotations.

Proposed Algorithm

Before introducing our privacy-preserving feature learning method, we discuss the privacy attack methods of the adversary. We then describe the proposed privacy-preserving and task-oriented representation learning.

Adversary: Learn to Invert

To design a proper defense mechanism, we first need to understand an adversary’s attack method. Specifically, we focus on the model inversion (MI) attack (Fredrikson, Jha, and Ristenpart 2015), where the goal of the adversary is to invert features back to an input. We further assume a blackbox attack, where the adversary has unlimited access to the models’ inference (Enc), but not the model parameters. This is extremely generous setting to the adversary since they can create a large-scale paired database of input $X \in X_2$ and the feature $Z = \text{Enc}(X)$. Here, we use $X_2$ to distinguish the adversary’s own dataset from the private training dataset $X_1$ of CSPs or their customers.

Given a paired dataset $\{(X \in X_2, Z)\}$, the adversary inverts the feature via a decoder $\text{Dec}^a : Z \rightarrow X$, which is trained to reconstruct the input $X$ from the feature $Z$ by minimizing the reconstruction loss$^3$:

$$L_p^a = \mathbb{E}_{((X \in X_2, Z))} \| \hat{X} - X \|^2,$$

where $X = \text{Dec}^a(Z)$. Note that the above does not involve backpropagation through an Enc. The inversion quality may be improved with the GAN loss (Goodfellow et al. 2014):

$$L_{\text{adv}}^a = \mathbb{E}_{Z} \log (1 - D(\hat{X})),$$

where $D$ is the discriminator, telling the generated images from the real ones by maximizing the following loss:

$$L_D^a = \mathbb{E}_{((X \in X_2, Z))} \left[ \log(1 - D(X)) + \log(D(\hat{X})) \right].$$

We can further improve the inversion quality by minimizing the perceptual distance (Dosovitskiy and Brox 2016a):

$$L_{\text{perc}}^a = \mathbb{E}_{((X \in X_2, Z))} \|g(\text{Dec}^a(Z)) - g(X)\|^2,$$

where we use the conv1 to conv5 layers of the VGG network (Simonyan and Zisserman 2015) pre-trained on the ImageNet for $g$. The overall training objective of an adversary thus can be written as:

$$\min_{\text{Dec}^a} \max_D \left\{ L_p^a + \mu_1 L_{\text{adv}}^a + \mu_2 L_{\text{perc}}^a \right\},$$

(5)

Protector: Learn “NOT” to Invert

Realizing the attack types, we are now ready to present the training objective of privacy-preserving and task-oriented representation. To learn a task-oriented representation, we adopt an MLP classifier $f$ that predicts the utility label $Y$ from $Z$ by minimizing the utility loss:

$$L_u = \mathbb{E}_{((X \in X_1, Y))} [L(f(Z), Y)],$$

$^3$Optimization-based inversion attack (Mohendran and Vedaldi 2015) may be considered instead, but it is not feasible in the black-box setting since the adversary has no access to Encs model parameters. Even in the white-box scenario, inversion by decoder may be preferred as it is cheaper to compute.
Table 1: Results on facial attribute prediction. $\lambda_2$, and $\mu_2$ are fixed to 0. We report the MCC averaged over 40 attributes as a utility metric and face and feature similarities as privacy metrics. The $\text{Enc}$ is trained using binary cross entropy (BCE) without or with the proposed adversarial reconstruction loss. Rows with $\dagger$ (#5 and #6) train $\text{Dec}^a$ with an extra data from MS-Celeb-1M, while those with $\ddagger$ (#7 and #8) use an extra data to train both $\text{Enc}$ and $\text{Dec}^a$. Different $\mu_1$'s in (4) are deployed for MI attack. #9 and #10 consider smiling attribute prediction for utility task and MCC is evaluated only for Smiling attribute.

<table>
<thead>
<tr>
<th>ID</th>
<th>$\lambda_1$</th>
<th>$\mu_1$</th>
<th>$\lambda_2$</th>
<th>$\mu_2$</th>
<th>Mean MCC $\dagger$</th>
<th>Face Sim. $\dagger$</th>
<th>Feature Sim. $\ddagger$</th>
<th>SSIM</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>$\mu_1$ = 0</td>
<td>0</td>
<td>$\mu_1$ = 0</td>
<td>0.641</td>
<td>0.551</td>
<td>0.835</td>
<td>0.231</td>
<td>13.738</td>
</tr>
<tr>
<td>2</td>
<td>$\lambda_1$ &gt; 0</td>
<td>$\mu_1$ = 0</td>
<td>$\lambda_2$ &gt; 0</td>
<td>$\mu_1$ = 0</td>
<td>0.612</td>
<td>0.515</td>
<td>0.574</td>
<td>0.221</td>
<td>13.423</td>
</tr>
<tr>
<td>3</td>
<td>$\lambda_1$ &gt; 0</td>
<td>$\mu_1$ = 0</td>
<td>$\lambda_2$ &gt; 0</td>
<td>$\mu_1$ = 0</td>
<td>0.641</td>
<td>0.585</td>
<td>0.835</td>
<td>0.240</td>
<td>14.065</td>
</tr>
<tr>
<td>4</td>
<td>$\lambda_1$ &gt; 0</td>
<td>$\mu_1$ &gt; 0</td>
<td>$\lambda_2$ &gt; 0</td>
<td>$\mu_1$ &gt; 0</td>
<td>0.612</td>
<td>0.513</td>
<td>0.574</td>
<td>0.277</td>
<td>13.803</td>
</tr>
</tbody>
</table>

With more data for training $\text{Dec}^a$ (ID #5 and #6) and both $\text{Enc}$ and $\text{Dec}^a$ (ID #7 and #8)

| 5  | $\lambda_1$ = 0 | $\mu_1$ = 0 | 0.641 | 0.594 | 0.864 | 0.250 | 14.132 |
| 6  | $\lambda_1$ > 0 | $\mu_1$ = 0 | 0.612 | 0.541 | 0.633 | 0.222 | 13.703 |
| 7  | $\lambda_1$ = 0 | $\mu_1$ = 0 | 0.651 | 0.579 | 0.834 | 0.263 | 14.432 |
| 8  | $\lambda_1$ > 0 | $\mu_1$ = 0 | 0.630 | 0.550 | 0.591 | 0.231 | 13.334 |

Single (Smiling) attribute prediction. MCC for Smiling attribute is reported in the parenthesis.

| 9  | $\lambda_1$ = 0 | $\mu_1$ > 0 | 0.001 (0.851) | 0.460 | 0.494 | 0.204 | 13.214 |
| 10 | $\lambda_1$ > 0 | $\mu_1$ > 0 | 0.044 (0.862) | 0.424 | 0.489 | 0.189 | 12.958 |

where $f(Z) = f(\text{Enc}(X))$, $Y$ is the ground-truth label for utility, and $L$ is the standard loss (e.g., cross-entropy) for utility. Note that $X_1$ is a private training data, which is not accessible to the adversary.

To learn a privacy-preserving feature against the MIA attack, an $\text{Enc}$ needs to output $Z$ that cannot be reconstructed into an input $X$ by any decoder. Unfortunately, enumerating all possible decoders is not feasible. Instead, we borrow the idea from adversarial learning (Goodfellow et al. 2014). To be more specific, the decoder ($\text{Dec}$) is trained to compete against an $\text{Enc}$ in a way that it learns to decode $Z$ of the current $\text{Enc}$ into $X$ by minimizing the reconstruction loss:

$$L_{p} = \mathbb{E}_{(X \in \mathcal{X}, Z)}[\|\text{Dec}(Z) - X\|^2].$$

(7)

In addition, one can improve the quality of reconstruction, resulting in a stronger adversary, using perceptual distance loss as in (4):

$$L_{perc} = \mathbb{E}_{(X \in \mathcal{X}, Z)}[\|g(\text{Dec}(Z)) - g(X)\|^2].$$

(8)

On the other hand, an $\text{Enc}$ aims to fool $\text{Dec}$ by maximizing the reconstruction loss or perceptual distance loss. Finally, the overall training objective of a protector is:

$$\min_{\text{Enc}, f} L_{u} - \lambda_1 L_{p} - \lambda_2 L_{perc},$$

(9)

while the decoder $\text{Dec}$ is updated by the same loss function as $\text{Dec}^a$ according to (5). Figure 2 shows the main modules of our method. We adopt alternative update strategy for the proposed learning framework. The $\text{Dec}^a$ and $D$ are updated first on the public data $X_2$ according to (5) while fixing the $\text{Enc}$ and $f$, and $\text{Enc}$ and $f$ are updated on the private data by (9) and so forth until convergence.

**Experimental Results**

We first introduce two types of privacy attack methods: the black-box MI attack and feature-level attack, and two metrics: face similarity and feature similarity. Next, we show the effectiveness of our method against privacy attacks in different scenarios.

**Dataset and Experimental Setup**

We use the widely-used CelebA (Liu et al. 2015) and MS-Celeb-1M (Guo et al. 2016) datasets for experiments. In most experiments, we split the CelebA dataset into three parts, $X_1$ with 160k images, $X_2$ with 40k images, and the test set $T$ with the rest. For ablation studies with different data for adversaries, we provide extra data for $X_2$ from the MS-Celeb-1M dataset. We use the ResNet-50 model (He et al. 2016) for feature representation and two fully connected layers for latent classifier $f$. The $\text{Dec}$ uses up-sampling layers to decode features to pixel images. We compare the proposed adversarial feature learning method to the baseline where $\text{Enc}$ is trained to minimize the binary cross entropy (BCE) for prediction for 40 binary facial attributes. More details regarding the networks can be found in the supplementary material.

**Evaluation against Data Privacy Attacks**

**Adversary’s Attacks.** Given an $\text{Enc}$, we evaluate the robustness of its feature by simulating the adversary’s attacks, i.e., the black-box MI attack and the feature-level attack.

While the MI attack aims at reconstructing the input to recover the holistic information, the feature-level attack aims to recover predefined private attributes from the feature. In other words, the adversary aims to learn a mapping function $M: Z \rightarrow C$ to reconstruct the feature of private attribute prediction network $C$ (e.g., face verification network):

$$\min_{M} \mathbb{E}_{X \in \mathcal{X}_2} [\|M(Z) - C(X)\|^2].$$

(10)

The privacy information is not well protected if one finds $M$ that successfully minimizes the loss in (10).
Utility Metric. We measure the attribute prediction performance of $f \circ \text{Enc}$ on the CelebA dataset, we use the Matthews correlation coefficient (MCC) (Boughorbel, Jarray, and El-Anbari 2017) as the evaluation metric:

$$\text{MCC} = \frac{(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where TP and FN stand for true positive and false negative, respectively. The MCC value falls into a range of $[-1, +1]$, in which +1, 0, and −1 indicate the perfect, random, and the worst predictions, respectively.

Privacy Metric. For the MI attack, we compare the reconstruction $\hat{X} = \text{Dec}^a(\text{Enc}(X))$ to $X$ by visual inspection and perceptual similarity. The face similarity between $X$ and $\hat{X}$ is computed by the cosine similarity of their deep features, e.g., layer 3 of $C$. Here, we use identity as a private attribute and an OpenFace face verification model (Amos, Ludwiczuk, and Satyanarayan 2016) for $C$. We also report the SSIM and PSNR of the reconstructed images to evaluate the realism. For the feature-level attack, we report the cosine similarity of features between $M(Z)$ and $C(X)$.

Empirical Results and Performance Analysis

Baseline. In Table 1, we first present main results by fixing $\mu_2$ and $\lambda_2$ equal to 0. The baseline model trained only with the binary cross entropy (ID #1, #3) of 40 attributes are compared with our model trained with negative reconstruction loss, i.e., $\lambda_1 = 1$ (ID #2, #4). We simulate the weak and strong adversaries without ($\mu_1 = 0$) or with ($\mu_1 > 0$) the GAN loss for training $\text{Dec}^a$. The proposed model maintains a reasonable task performance with a decrease in face similarity and feature similarity (#1 vs #2, #3 vs #4 in Table 1). Such improved privacy protection is also shown in face or feature similarities and visual inspection (Figure 3).

More Data. We validate the proposed model by adding more data (10k images from MS-Celeb-1M) to $X$ (#5 and #6) to train $\text{Dec}^a$. While extra data similar to the data distribution of $X$ helps the adversary to improve $\text{Dec}^a$, as shown from the face similarity comparisons between #1 and #5 (0.851 to 0.865), the $\text{Enc}$ trained with our method is still more robust to the MI attack. Furthermore, providing the same amount of data for the $\text{Enc}$ training (#7 and #8) improves the MCC slightly, while both face similarity and feature similarity decrease more. This demonstrates that more data for the $\text{Enc}$ training will help protect the privacy. The reason is more data would help better simulate a strong fake adversary during the training stage, such that the $\text{Enc}$ can learn to encode images into more secure features.

Single Attribute. We consider a single attribute prediction. Compared to 40 attributes prediction, the network needs to maintain much less amount of information for a single attribute prediction, and thus it should be easier to make
Table 2: Results with different $\lambda_2$ in the training stage. Other hyperparameters are fixed: $\lambda_1 = 1, \mu_1 = 0, \mu_2 = 1$.

<table>
<thead>
<tr>
<th>$\lambda_2$</th>
<th>MCC ↑</th>
<th>Face Sim. ↓</th>
<th>Feat. Sim. ↓</th>
<th>SSIM</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.631</td>
<td>0.631</td>
<td>0.862</td>
<td>0.300</td>
<td>15.445</td>
</tr>
<tr>
<td>1</td>
<td>0.582</td>
<td>0.575</td>
<td>0.710</td>
<td>0.299</td>
<td>15.371</td>
</tr>
<tr>
<td>2</td>
<td>0.455</td>
<td>0.545</td>
<td>0.604</td>
<td>0.273</td>
<td>14.920</td>
</tr>
<tr>
<td>3</td>
<td>0.417</td>
<td>0.528</td>
<td>0.568</td>
<td>0.248</td>
<td>14.454</td>
</tr>
<tr>
<td>4</td>
<td>0.311</td>
<td>0.507</td>
<td>0.542</td>
<td>0.225</td>
<td>14.048</td>
</tr>
<tr>
<td>5</td>
<td>0.255</td>
<td>0.502</td>
<td>0.530</td>
<td>0.224</td>
<td>14.047</td>
</tr>
<tr>
<td>10</td>
<td>0.000</td>
<td>0.294</td>
<td>0.374</td>
<td>0.158</td>
<td>12.899</td>
</tr>
</tbody>
</table>

features privacy-protected. For example, the baseline $\text{Enc}$ trained for a single attribute prediction (#9) has significantly lower face and feature similarities compared to that trained for 40 attributes prediction (#1). Our method (#10) further improves the privacy-protection thanks to the proposed negative reconstruction loss. As shown in the rightmost column of Figure 3, the reconstructed images look like the mean faces reflecting only the smiling attribute, as the information of all other attributes is automatically discarded.

**Visualization of Latent Representations**

We visualize the learned features using t-SNE (Maaten and Hinton 2008) to analyze the effectiveness of the proposed method. Figure 4 shows the t-SNE plots for two scenarios, either trained for 40 attribute prediction tasks or a single (Smiling) attribute prediction task. For presentation clarity, we colorcode data points with respect to two attributes, namely, Eyeglasses and Smiling.

The baseline model features are well separable along both attributes when trained for 40 attributes. On the other hand, the features of the model trained with adversarial reconstruction loss are more uniformly distributed along these two attributes since the encoder is forced to discard the private information relevant to the reconstruction as much as possible.

For the single attribute case, only the Smiling attribute is preserved for both models. The red and blue dots are mixed in the right side of Figure 4, which indicates the attribute Eyeglasses is recognizable or separable. That is because the attribute Eyeglasses is not the target training utility. The adversarial reconstruction loss does not turn out to be particularly effective since the retained information in the feature space is already minimal but sufficient for a single attribute prediction and adding an extra loss to interfere the reconstruction does not have a large benefit. This also corresponds to the last columns of Figure 3, where the reconstructed images are almost mean faces with only smiling attributes preserved.

**Ablation Study of Perceptual Distance Loss**

We analyze the influence of the perceptual distance loss on the privacy preserving mechanism. To pinpoint its impact, we perform experiments by changing $\lambda_2$, which corresponds to the weight of the perceptual distance term in (8), while fixing other hyperparameters for $\text{Enc}$ ($\lambda_1 = 1$) and $\text{Dec}^a$ ($\mu_1 = 0, \mu_2 = 1$). All utility and privacy metrics are reported in Table 2 and the reconstructed images are presented in Figure 5.

Changing $\lambda_2$ clearly demonstrates the trade-off between utility and privacy. As we increase $\lambda_2$, the model becomes more privacy-protecting as demonstrated by the lower face and feature similarities, but this comes at the cost of decreased utility performance (mean MCC). Our quantitative analysis is also consistent with the visual results, where the reconstructed images in Figure 5 contain less sensitive information as $\lambda_2$ increases and tend to be a mean face when $\lambda_2$ becomes as high as 10. The trade-off between utility and privacy suggests that we can achieve different level of privacy in real applications.

**Ablation Study of Different Layers**

As in (Dosovitskiy and Brox 2016b; 2016a), the MI attack becomes much easier when low- or mid-level features are used for inversion. We analyze the effect of features from different layers on the utility and privacy. We choose six different layers of intermediate features in $\text{Enc}$, and adapt the network structures of $\text{Dec}$ and $f$ accordingly. Specifically, $\text{Enc}$ and $\text{Dec}$ are symmetric to each other, and $f \circ \text{Enc}$ is the entire ResNet-50 architecture followed by 2 fully connected layers at all cases.

We first compare the face similarity and MCC in Table 3. The face similarity decreases as the layer goes deeper for the baseline model because the information becomes abstract during the forwarding process. With adversarial training, the face similarity are reduced compared to those without adversarial training, while the MCC is not affected significantly. Furthermore, the face similarity with adversarial training in
lower layers is generally lower than that in deeper layers.

Next, we present within-class variance $S_w$, between-class variance $S_b$ and the LDA score $S_b/S_w$. A low LDA score indicates that the relative distance among the features of different identities is small, thus more privacy-preserving, whereas the distance between two features of the same identity becomes large. As shown in Table 3, the LDA score of baseline model decreases as the layer goes deeper. In addition, the LDA score of our model is generally smaller than that of the baseline, which further validates that the features with adversarial training are more uniformly distributed.

**Relation to Information Bottleneck**

While our method is developed by integrating potential attacks from adversary (e.g., decoding latent representation into the pixel space), our method can be understood from the information-theoretic perspective. The objective function can be mathematically formalized using mutual information and conditional entropy:

$$\min_{\text{Enc}} \max_{\text{Dec}} I(\text{Dec}(Z); X) + H(Y|f(Z)), \quad (11)$$

where $Z = \text{Enc}(X)$. Note that our objective resembles that of information bottleneck methods (Tishby, Pereira, and Bialek 2000; Achille and Soatto 2018; Alemi et al. 2017) except that we introduce the decoder to estimate mutual information via a min-max game between Enc and Dec (Belghazi et al. 2018). The mutual information term can be reduced as:

$$I(\text{Dec}(Z); X) = H(X) - H(X|\text{Dec}(Z)). \quad (12)$$

Since $H(X)$ is a constant, the protector’s objective is:

$$\min_{\text{Enc}} \max_{\text{Dec}} -H(X|\text{Dec}(Z)) + H(Y|f(Z)). \quad (13)$$

If we use the reconstruction error for the first term:

$$H(X|\text{Dec}(Z)) = \mathbb{E}_{\{X,Z\}}[||\text{Dec}(Z) - X||^2], \quad (14)$$

(13) is realized as a minimization of negative reconstruction loss for updating Enc. If we use reconstruction error and perceptual error for the first term:

$$H(X|\text{Dec}(Z)) = \lambda_1 \mathbb{E}_{\{X,Z\}}[||\text{Dec}(Z) - X||^2] + \lambda_2 ||g(\text{Dec}(Z)) - g(X)||^2, \quad (15)$$

(13) is realized as a minimization of negative reconstruction loss and negative perceptual loss for updating Enc as in (9).

Aside from empirical results, the information-theoretic perspective of our method also reveals some limitations of the current approach. First, maximizing $H(X|\text{Dec}(Z))$ may not directly provide privacy-preservation to latent representations, which is evident from the data-processing inequality $H(X|\text{Dec}(Z)) \geq H(X|Z)$. To protect privacy against universal attacks, it is necessary to develop methods to maximize $H(X|Z)$. Furthermore, our algorithm can defend against data reconstruction adversary, but there might be useful private information remaining that adversaries can take advantage of, such as the face ethnicity information from the skin color of reconstructed images, even though the reconstructed images are not recognized the same as the input images. However, the proposed method provides an effective method to protecting against private attributes, which is of great interest towards learning explainable representations in the future.

**Conclusions**

We propose an adversarial learning framework to learn a latent representation that preserves visual data privacy. Our method is developed by simulating the adversary’s expected behavior for the model inversion attack and is realized by alternating update of Enc and Dec networks. We introduce quantitative evaluation methods and provide comprehensive analysis of our adversarial learning method. Experimental results demonstrate that our algorithm can learn privacy-preserving and task-oriented representations.

**Acknowledgements.** This work is supported in part by the NSF CAREER Grant #1149783 and gifts from NEC.

**References**


Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep variational information bottleneck. *ICLR*.


---

Table 3: Evaluation on different layers. We show that our method “w/ adv” maintains a good mean MCC while reducing the face similarity to protect the privacy, consistently improving across all feature layers compared to “w/o adv”. Furthermore, our method shows smaller LDA scores, which indicate that the relative distance among the features of different identities becomes smaller, hence benefiting privacy preservation.

<table>
<thead>
<tr>
<th></th>
<th>Conv1</th>
<th>Conv2</th>
<th>Conv3</th>
<th>Conv4</th>
<th>Avg Pool</th>
<th>FC Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>w/o adv</td>
<td>w/ adv</td>
<td>w/o adv</td>
<td>w/ adv</td>
<td>w/o adv</td>
<td>w/ adv</td>
</tr>
<tr>
<td>Face Sim ↓</td>
<td>0.98</td>
<td>0.36</td>
<td>0.95</td>
<td>0.48</td>
<td>0.74</td>
<td>0.49</td>
</tr>
<tr>
<td>Mean MCC †</td>
<td>0.64</td>
<td>0.65</td>
<td>0.65</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>Within Var ($S_w$)</td>
<td>1468.77</td>
<td>768.29</td>
<td>1543.09</td>
<td>1523.31</td>
<td>2067.54</td>
<td>1779.75</td>
</tr>
<tr>
<td>Between Var ($S_b$)</td>
<td>2884.42</td>
<td>2188.88</td>
<td>2726.15</td>
<td>597.48</td>
<td>1608.97</td>
<td>598.39</td>
</tr>
<tr>
<td>LDA Score ↓</td>
<td>1.96</td>
<td>0.28</td>
<td>1.77</td>
<td>0.39</td>
<td>0.78</td>
<td>0.34</td>
</tr>
</tbody>
</table>


Appendix

Network Architectures

The encoder is set to be the standard ResNet-50 architecture. We refer an official implementation for a detailed reference\(^4\). Unless otherwise stated, we extract the feature for Enc from the fc layer. The architecture of the decoder is almost reverse to the Enc with up-sampling, as shown in Figure 6.

In the ablation study with different layers of features, we choose six different layers of ResNet-50. The architectures of Dec and latent classifier \(f\) change accordingly. For instance, if we choose layer3 as the intermediate layer, then the decoder should start from layer3 to layer1 and the latent classifier should start from layer4 to fc with two additional fully connected layers followed.

Visualization with Perceptual Distance Loss

In Figure 7, we visualize additional results for reconstruction to demonstrate the trade-off between utility and privacy by changing the regularization coefficient of perceptual distance loss \(\lambda_2\) while fixing other hyperparameters for Enc (\(\lambda_1 = 1\)) and Dec\(^a\) (\(\mu_1 = 0, \mu_2 = 1\)).

\(^4\)https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py

---

Figure 6: Dec and Dec\(^a\) network architectures.
Figure 7: Reconstructed images with different $\lambda_2$ (shown on top of each image) by fixing $\lambda_1 = 1$, and $\mu_1 = 0, \mu_2 = 1$. 