COMPRESSED DATA SHARING BASED ON INFORMATION BOTTLENECK MODEL

Behrooz Razeghi*‡, Shideh Rezaeifar*‡, Sohrab Ferdowsi†, Taras Holotyak*, Slava Voloshynovskiy*
* University of Geneva † HES-SO Geneva

***Abstract*—In this paper, we consider privacy-preserving compressed image sharing, where the goal is to release compressed data whilst satisfying some privacy/secrecy constraints yet ensuring image reconstruction with a defined fidelity. The privacy-preserving compressed image sharing is addressed using a machine learning framework based on an information bottleneck with a shared secret key for authorized users. In contrast, an adversary observing the protected compressed representation tries to either reconstruct the data or deduce some privacy-sensitive attributes such as gender, age, etc. The inference task on the adversary's side is performed without the knowledge of the shared secret key and is based on an adversarial mutual information maximization between the privacy-protected compressed representation and targeted attributes. The proposed framework is experimentally validated on the CelebA dataset.**

I. Introduction

In this work, we address the problem of privacy-preserving high-dimensional compressed image sharing, where the goal is to release compressed data while also satisfying some privacy/secrecy constraints. In the era of big data, efficient yet secure data communication and storage among different parties have become a necessity. On the other hand, big data and considerable means of data collection provide abundant knowledge and rich tools for adversaries to launch various attacks. Hence, the challenge is to design a practical low-complexity compressed data sharing mechanism that satisfies privacy constraints. To alleviate this problem, we propose an efficient information-preserving compression mechanism based on (1) dimensionality reduction, (2) sparsification, (3) quantization and (4) obfuscation. The proposed mechanism is inspired by Shannon's notion of information-theoretic secrecy and is based on the Information Bottleneck (IB) principle [1].

Given two correlated data random variables $(\mathbf{C}, \mathbf{X})$, the goal of original IB model [1] is to find a compressed representation $\mathbf{Z}$ via a stochastic map $p(\mathbf{z} \mid \mathbf{x})$ such that: $\mathbf{C} \rightarrow \mathbf{X} \rightarrow \mathbf{Z}$ forms a Markov chain, and the representation $\mathbf{Z}$ preserves all relevant information in $\mathbf{X}$ about desired variable $\mathbf{C}$. Given the mutual information as both a cost function and a regularizer, an optimal representation $\mathbf{Z}$ satisfying a certain compression-relevance trade-off constraint is then found by minimizing the Lagrangian functional $\mathcal{L}_{IB} = I(\mathbf{C};\mathbf{Z}) - \beta I(\mathbf{X};\mathbf{Z})$, where $I(\mathbf{C};\mathbf{Z})$ and $I(\mathbf{X};\mathbf{Z})$ are Shannon's mutual information expressions between the corresponding variables, and β is the Lagrangian multiplier. Inspired by the original formulation of IB method [1], abundant characterizations, generalizations and applications have been proposed [2]–[13]. According to [14], one can formulate the reconstruction task, i.e., consider an unsupervised IB model, as $\mathbf{C} \equiv \mathbf{X}$. This interpretation yields a more general form of the variational auto-encoder [15].

‡ The authors contributed equally.
Full version is available at: https://github.com/BehroozRazeghi/.

A. Contribution

The two main contributions of our approach are as follows.

a) Information Bottleneck Model with Shared Secrecy: We propose a new formulation of Information Bottleneck with 'shared secrecy' between the encoder-decoder pair of *defender*. To the best of our knowledge, this is the first time such a problem is addressed in the machine learning setup with information-theoretic secrecy. The proposed mechanism can be utilized in data sharing via "noisy" channels, where the celebrated Shannon's principle of compression-encryption-decryption-decompression does not apply directly due to the noisy nature of communications. Although in this paper we primarily focus on the passive adversary, at the same time, we envision an active adversary scenario in future when the adversary can modify the compressed representation trying to trick the decoding. This scenario might be considered as a sort of adversarial attack in the latent compressed space.

b) New Insights for Secure Machine Mechanism Design: Nowadays, the security of modern machine learning methods is generally addressed either using the training data advantage of the defender over the adversary or various distributed architectures to cope with the privacy of data and to limit the access of adversary to the same training data [16]–[18]. Moreover, the addition of various ambiguation factors preventing data recovery or re-identification has been suggested [19].

II. Proposed privacy-preserving compression

A. General Setup

Given the high-dimensional observed data $\mathbf{X}$, the defender intends to release a compact public representation $\mathbf{Z}_{\mathrm{p}}$ to provide some utility service (e.g., recognition, reconstruction, etc.) for some authorized parties. However, the defender is concerned with the risk that the original (private) data $\mathbf{X}$ might get exposed to an adversary eavesdropping on the released representation. To preserve privacy, the defender applies an obfuscation mechanism to the original compressed data before releasing it. This data-driven obfuscation mechanism smoothly trades off the informativeness of the bottleneck latent representation for the utility service at hand against the compressiveness of the bottleneck variable from original (private) data, while at the same time minimizes the privacy leakage.

B. Threat Model

We consider two threat models. The first model addresses the adversarial reconstruction from the publicly released compressed representation $\mathbf{Z}_{\mathrm{p}}$ under the unknown secret key used for the ambiguation. It should be pointed out that to be compliant with Kerkhoff's principle, we assume that the attacker knows the algorithm of encoding, but the only unknown parameter is a secret key that is at the same time independent of

the data. Therefore, given the training set of publicly released vectors protected by any random key, the adversary targets to design a decoder that would map the public vector $\mathbf{Z}_\mathrm{p}$ into the reconstructed image $\mathbf{X}_\mathrm{a}$ representing the variable $\mathbf{Y}$ under the reconstruction attack that at the same time represents a regression problem. This type of attack can be linked to a *model inversion attack* [20].

The attribute inference attack refers to a classification problem when the adversary aims at *inferring* some sensitive attribute of data, e.g, gender or age, rather than to *reconstruct* the raw data itself [21]–[25]. To this end, we suppose that the adversary observes the compressed and protected public data $\mathbf{Z}_\mathrm{p}$ and tries to infer one of the above attributes. It is achieved by training a classifier or decoder on a training set of public representations $\mathbf{z}_\mathrm{p}$ and available attributes $\mathbf{c}$. As in the previous case, we assume that the attacker knows the mechanism of data encoding and ambiguiation, but the secret key used for the ambiguiation remains unknown.

C. Problem Formulation

Given the observed data $\mathbf{X}$ the objective is to find the private bottleneck representation $\mathbf{Z}$ and its public counterpart $\mathbf{Z}_\mathrm{p}$ such that given another random variable $\mathbf{K}$, playing a role of a secret key available for the authorized parties only, we have $\Pr\{\mathbf{Z}_\mathrm{p} \mid \mathbf{K}\} = \Pr\{\mathbf{Z}\}$. The random variable $\mathbf{Z}$ denotes the clean representation and $\mathbf{Z}_\mathrm{p}$ denotes the released (public) representation. One can interpret $\mathbf{Z}$, $\mathbf{Z}_\mathrm{p}$, and $\mathbf{K}$ as plain-text, cipher-text, and secret key, respectively. Accordingly, the problem can be formulated by minimization of Lagrangian:

$$\mathcal{L}^\mathrm{d} = I(\mathbf{X};\mathbf{Z}_\mathrm{p} \mid \mathbf{K}) - \beta I(\mathbf{X};\mathbf{Z}_\mathrm{p} \mid \mathbf{K}), \tag{1}$$

on the side of defender. Note that the given key $\mathbf{K}$ is independent of $\mathbf{X}$ due to the security reasons, the Lagrangian functional reduces to $I(\mathbf{X};\mathbf{Z}) - \beta I(\mathbf{X};\mathbf{Z})$.

D. Practical Design of Defender's Mechanism

Let $q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})$ denotes the parametrized stochastic mapping corresponding to $p(\mathbf{z} \mid \mathbf{x})$, and let $p_{\mathcal{D}}(\mathbf{x})$ denotes the empirical data distribution. In this case, $q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z}) = p_{\mathcal{D}}(\mathbf{x}) q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ denotes the joint data distribution. Also, let $q_{\boldsymbol{\phi}}(\mathbf{z})$ denotes the aggregated distribution of latent space. The mutual information $I(\mathbf{X};\mathbf{Z})$ can be written as:

$$I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z})}\left[\log \frac{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}{q_{\boldsymbol{\phi}}(\mathbf{z})}\right] = H_{\boldsymbol{\phi}}(\mathbf{Z}) - H_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X}), \tag{2}$$

where $H_{\boldsymbol{\phi}}(\mathbf{Z}) = -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z})}[\log q_{\boldsymbol{\phi}}(\mathbf{z})]$, and $H_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X}) = -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z})}\left[\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right]$. Since entropy $H_{\boldsymbol{\phi}}(\mathbf{Z})$ requires computation of marginal distribution $q_{\boldsymbol{\phi}}(\mathbf{z}) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{X})}[q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})]$ that is computationally expensive, we will proceed with the variational approximation of $q_{\boldsymbol{\phi}}(\mathbf{z})$ by a distribution $p_{\boldsymbol{\theta}}(\mathbf{z})$.

Therefore, we decompose $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z})$ as follows:

$$I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) = \underbrace{\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}=\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{z}_p)\right)\right]}_{\mathrm{A}} - \underbrace{D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z})\right)}_{\mathrm{B}}, \tag{3}$$

where the term (A) denotes the KL-divergence $\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}=\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{z})\right)\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{x})}[\log \frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z})}] = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log \frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z})}]\right]$ and the term (B) denotes the KL-divergence $D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z})\right) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{x})}[\log \frac{q_{\boldsymbol{\phi}}(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{z})}] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z})}[\log \frac{q_{\boldsymbol{\phi}}(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{z})}]$.

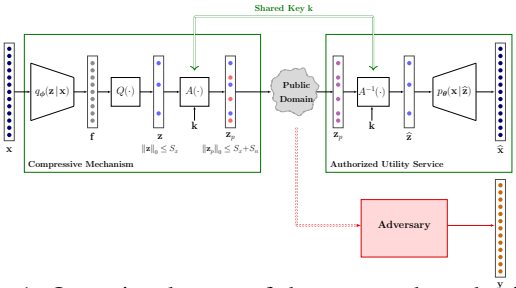


Fig. 1: Operational setup of the proposed mechanism.

a) Reconstruction Task: : Given the key $\mathbf{K}$, the second term in Lagrangian can be decomposed as $I(\mathbf{X};\mathbf{Z}) = \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{p(\mathbf{x}|\mathbf{z})}{p_{\mathcal{D}}(\mathbf{x})}\right]$. This mutual information can be lower bounded using a variational lower bounded as $I(\mathbf{X};\mathbf{Z}) \geq I_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) = H_{\mathcal{D}}(\mathbf{X}) - H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{X}|\mathbf{Z})$ and simultaneously $I_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}) = H(p_{\mathcal{D}}(\mathbf{x});p_{\boldsymbol{\theta}}(\mathbf{x})) - D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{x})\right) - H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{X}|\mathbf{Z})$. Since $H(p_{\mathcal{D}}(\mathbf{x});p_{\boldsymbol{\theta}}(\mathbf{x})) \geq 0$, one can consider lower bound $I^L_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}) = -D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{x})\right) - H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{X}|\mathbf{Z})$. Finally, if the utility service is a reconstruction task, the *defender* minimization Lagrangian functional can be written as:

$$\mathcal{L}^\mathrm{d}(\boldsymbol{\theta},\boldsymbol{\phi}) = I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) - \beta I^L_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}), \tag{4}$$

leading to the minimization problem:

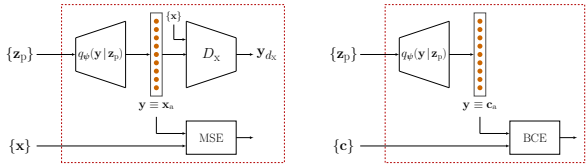
$$\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\phi}}\right) = \arg\min_{(\boldsymbol{\theta},\boldsymbol{\phi})} \mathcal{L}^\mathrm{d}(\boldsymbol{\theta},\boldsymbol{\phi}). \tag{5}$$

b) Classification Task: If the utility service of the defender targets some attributes denoted by $\mathbf{C}$, then the second term of IB Lagrangian can be formulated as:

$$I^\mathrm{S}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{C}) \triangleq H(p(\mathbf{c});p_{\boldsymbol{\theta}}(\mathbf{c})) - D_{\mathrm{KL}}(p(\mathbf{c}) \| p_{\boldsymbol{\theta}}(\mathbf{c})) - H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{C}|\mathbf{Z}), \tag{6}$$

with $H(p(\mathbf{c});p_{\boldsymbol{\theta}}(\mathbf{c})) = -\mathbb{E}_{p(\mathbf{c})}[\log p_{\boldsymbol{\theta}}(\mathbf{c})]$ denoting a cross-entropy between $p(\mathbf{c})$ and $p_{\boldsymbol{\theta}}(\mathbf{c})$, and $D_{\mathrm{KL}}\left(p(\mathbf{c}) \| p_{\boldsymbol{\theta}}(\mathbf{c})\right) = \mathbb{E}_{p(\mathbf{c})}[\log \frac{p(\mathbf{c})}{p_{\boldsymbol{\theta}}(\mathbf{c})}]$ to be a KL-divergence between the prior class label distribution $p(\mathbf{c})$ and the estimated one $p_{\boldsymbol{\theta}}(\mathbf{c})$. One can assume different forms of encoding of labels $\mathbf{c}$ but one of the most often used forms is *one hot label encoding* that leads to the categorical distribution of priors $p(\mathbf{c}) = \mathrm{cat}(\mathbf{c})$. It also naturally imposes a sparsity constraint such that $\|\mathbf{c}\|_0 = 1$, where $\|.\|_0$ denotes the ℓ_0-"norm". Since $H(p(\mathbf{c});p_{\boldsymbol{\theta}}(\mathbf{c})) \geq 0$ and $D_{\mathrm{KL}}\left(p(\mathbf{c}) \| p_{\boldsymbol{\theta}}(\mathbf{c})\right) \geq 0$, it leads to the lower bound $I^S_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{C}) \geq I^{\mathrm{S}_L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{C})$, with $I^{\mathrm{S}_L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{C}) \triangleq -H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{C}|\mathbf{Z})$. Accordingly, one can reformulate the supervised Lagrangian functional as: $\mathcal{L}^{\mathrm{S}_L}(\boldsymbol{\phi},\boldsymbol{\theta}) \propto I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) - \beta I^{\mathrm{S}_L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{C}) = I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) + \beta H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{C}|\mathbf{Z})$. In this paper, we only consider the reconstruction problem for the defender. Accordingly we will consider the mean square error (MSE) reconstruction counterpart of conditional entropy $H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{X}|\mathbf{Z})$ in (**??**) leaving the classification problem for the future research.

c) Ambiguation Mechanism: Given the data samples $\mathbf{x}$, the encoder generates a sparse representation $\mathbf{z}$ such that $\|\mathbf{z}\|_0 \leq S_x$. The ambiguation mechanism adds S_n random components, with the same statistics as sparse code, to the



(a) (b)

Fig. 2: Adversary Model: (a) reconstruction task ($\mathbf{Y} \equiv \mathbf{X}_\mathrm{a}$), (b) classification task ($\mathbf{Y} \equiv \mathbf{C}_\mathrm{a}$).

orthogonal complement of sparse representations produced by the encoder. Note that this mechanism design provides the condition $\Pr\{\mathbf{Z}_\mathrm{p} \mid \mathbf{K}\} = \Pr\{\mathbf{Z}\}$ introduced in Section II-C. The general block-diagram of our model is depicted in Fig. 1.

E. Adversarial Decoding Strategies

We consider two adversarial strategies attacking the privacy of data along with the secure compressed data sharing problem. The first strategy targets unauthorized reconstruction and is schematically shown in Fig. 2a. The second strategy focuses on revealing privacy-sensitive attributes such as gender or age and is shown in Fig. 2b. We will refer to this strategy as adversarial classification or attribute inference attack. For both strategies, we assume that the adversary has access to some training data represented by the compressed public data and either original data for the reconstruction strategy or the class labels for the classification strategy.

In the case of reconstruction attack, the targeted attribute $\mathbf{Y}$ corresponds to the original data/image $\mathbf{X}_\mathrm{a}$. Therefore, the adversary wishes to maximize the mutual information:

$$\mathcal{L}^\mathrm{a}(\boldsymbol{\psi}) = \mathbb{E}_{p(\mathbf{k})}\left[I^L_{\boldsymbol{\psi}}(\mathbf{Z}_\mathrm{p}; \mathbf{X} \mid \mathbf{K})\right], \tag{7}$$

under a set of unknown keys generated from some assumed distribution $p(\mathbf{k})$ that also leads to the maximization problem $\widehat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} \mathcal{L}^\mathrm{a}(\boldsymbol{\psi})$. Practically, the attacker tries to design an adversarial parametrized decoder $q_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{z}_\mathrm{p})$ with $\mathbf{y} = \mathbf{x}_\mathrm{a}$ by providing the closes reconstruction to the original data $\mathbf{x}$.

The above mutual information can be lower bounded by the adversarially trained discriminator and MSE term, corresponding to the conditional entropy, under the unknown key corresponding to $-D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x}) \| p_{\boldsymbol{\psi}}(\mathbf{x}|\mathbf{k})\right) - H_{\boldsymbol{\psi}}(\mathbf{X}|\mathbf{Z}, \mathbf{K}))$ as shown in Fig. 2a. It is important to point out that in the considered interpretation, the adversarial decoder treats the key as a conditional random factor and tries to optimize the decoding on average.

In the case of attribute inference strategy, the targeted attribute $\mathbf{Y}$ corresponds to some data attributes $\mathbf{C}_\mathrm{a}$ representing some classes as shown in Fig. 2b. The corresponding adversarial objective is: $\mathcal{L}^\mathrm{a}(\boldsymbol{\psi}) = \mathbb{E}_{p(\mathbf{k})}\left[I^L_{\boldsymbol{\psi}}(\mathbf{Z}_\mathrm{p}; \mathbf{C} \mid \mathbf{K})\right]$. It can be lower bounded by the negative cross-entropy $-H_{\boldsymbol{\psi}}(\mathbf{C}|\mathbf{Z}_\mathrm{p}, \mathbf{K})$.

Therefore, under both considered adversarial strategies, the adversary trains the decoder $\boldsymbol{\psi}$ to reconstruct under *unknown* key[1]. We aim at investigating whether our secure compressive mechanism can withstand the considered attacks. It should be pointed out that we do not perform any special optimization to withstand the adversarial attribute inference attacks and consider the protection as a byproduct of secure compression. Therefore, the proposed scheme should not be expected to be protected from this type of attacks.

[1]Multiple realizations are utilized during adversarial training

III. Experiments

In this section, we quantitatively validate the proposed framework under both considered adversarial strategies. This section is structured as follows. First, we will introduce the experimental settings including the dataset and the attacker/defender setup and their corresponding assumptions. Then, we demonstrate our main results for two different attack strategies, namely, reconstruction and attribute inference attacks.

A. Experimental Setup

a) Dataset: We conducted experiments on the celebrity face image dataset, CelebA [26], which consists of over 200,000 celebrity images, where each image has been annotated with 40 different attributes. Every input image is center-cropped by 178x178 and then resized to 128x128. We randomly split the dataset and picked 80% of the dataset for training the network and the remainder for testing.

b) Defender setup: The defender's objective is twofold. On one hand, s/he wants to provide maximal utility from the data to the authorized users. On the other hand, the chances of the attacker to infer the privacy-sensitive content should be minimized. This trade-off is best achieved when, firstly, the latent representation $\mathbf{Z}$ has a carefully optimized rate-distortion profile. Secondly, there is to be an efficient mechanism to ambiguate and disambiguate the latent representations.

Therefore, it makes sense for the defender to use a sparse autoencoder, where compact latent codes provide a useful representation of the data. At the same time, the latent code should be sufficiently sparse so that ambiguazation noise can be injected to the zero-values of the representation making the attacker's inference unsuccessful. Note that if the rate-distortion optimization of the attacker fails, the latent codes will not be evenly spread out across the space, increasing the chances of unauthorized inference.

Here we use typical reconstruction losses for distortion minimization of the autoencoder during training of the defender, while we optimize the rate by construction, i.e., we impose a prescribed S_x-sparsity to $\mathbf{Z}$. In order to impose such a sparsity constraint, we use the top-S_x operator as the non-linearity before the latent layer, where the S_x values with the largest magnitudes are retained, otherwise zeroed-out. Since the derivative of this function exists and is non-zero at most of its operational range, it passes healthy gradients and does not show slow-down in training.

c) Attacker setup: The adversary has a set of publicly available protected representations with the corresponding labels, and his goal is to train the decoder to either reconstruct the original image or infer some sensitive attributes from these images, namely gender or age. The former is categorized in *model inversion attack* [20] and the latter in *attribute inference attack* [21] [22].

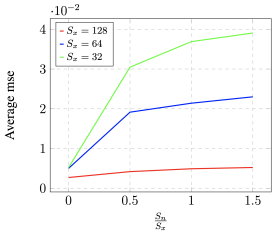


Fig. 3: Reconstruction attack performance in terms of reconstruction MSE. The adversarial decoder is trained for different sparsity levels and different ambiguation ratios.

Attribute	Sparsity	$\frac{S_n}{S_x}$ 0	0.5	1	1.5
Gender	128	94.69	93.72	93.60	93.52
Gender	64	94.35	89.68	88.78	86.78
Gender	32	93.26	64.61	62.22	58.71
Age	128	85.028	84.46	83.20	82.98
Age	64	83.75	82.15	81.61	81.17
Age	32	83.08	78.44	77.48	77.42

TABLE I: Attribute inference attack performance in terms of classification accuracy for CelebA dataset for different sparsity levels and different ambiguation ratios.

B. Reconstruction attack

The adversary model for reconstruction attack, as shown in Fig. 2a. consists of a decoder $q_{\boldsymbol{\psi}}(\mathbf{y} = \mathbf{x}_\mathrm{a}|\mathbf{z}_\mathrm{p})$ and a discriminator D_x trained on the original data and the reconstructed data $\mathbf{x}_\mathrm{a} = \mathrm{Decoder}(\mathbf{z}_\mathrm{p})$. The decoder network maps the protected noisy representation $\mathbf{z}_\mathrm{p}$ to the data space. The discriminator assigns the probability $\mathbf{y}_{d_\mathrm{x}} = D_\mathrm{x}(\mathbf{x}) \in [0, 1]$, where $\mathbf{x}$ is an actual training sample and the probability $1 - \mathbf{y}_{d_\mathrm{x}}$, where $\mathbf{x}$ is generated by the decoder network $q_{\boldsymbol{\psi}}(\mathbf{x}_\mathrm{a}|\mathbf{z}_\mathrm{p})$.

The discriminator is trained to find a binary classifier, which gives the best possible discrimination between the true and reconstructed data $\mathcal{L}_{D_\mathrm{x}} = \log(D_\mathrm{x}(\mathbf{x})) + \log(1 - D_\mathrm{x}(\mathbf{x}_\mathrm{a}))$.

Simultaneously, the decoder network is trained to fool the discriminator and minimize the MSE loss between reconstructed and original images. The detailed architecture of the decoder and discriminator models are reported in the full version. We conducted experiments under the reconstruction attack for different sparsities S_x, and different ambiguation ratios $\frac{S_n}{S_x}$. The performance of the attack model is evaluated using the average MSE distance between the input image and the reconstructed one, as reported in Fig. 3. The visualization of reconstructed images by the adversary network is shown in Fig. 4. The results show that for high S_x, the adversary can reconstruct the images with a reasonable fidelity. More specifically, for $S_x = 128$, neither visual quality nor reconstruction error has changed drastically with the increase of ambiguization noise. However, for the lower values of S_x equal to 64 and 32, the reconstruction error remarkably increases for higher ambiguization ratios. Hence, control of the sparsity allows the designer to ensure that the achievable adversary reconstruction is visually poor and the adversary fails with adversarial reconstruction under the proposed secure compression.

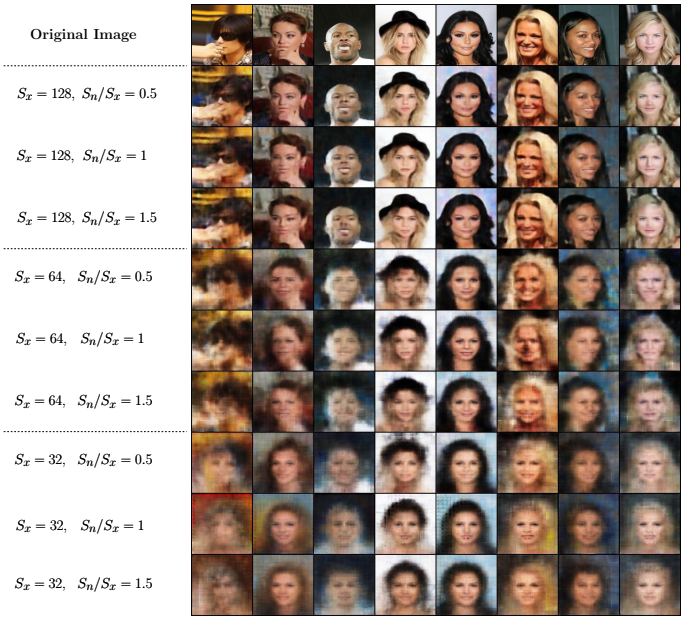


Fig. 4: Visual performance of reconstruction attack.

C. Attribute inference attack

For the attribute inference attack, we adopted Resnet50 [27] to map the protected representation, $\mathbf{z}_\mathrm{p}$ to the predicted attribute. The detailed architecture of the classifier is reported in full version. CelebA face dataset [26] includes 40 binary facial attributes, among which we have considered gender (male/female) and age (young/old) in our attack model. The performance of attack model for different sparsities, S_x, and different ambiguation ratios, $\frac{S_n}{S_x}$ is reported in Table. I.

The results show that the proposed compression is not entirely protected against leaking binary attributes as it was initially expected. This issue is easily explained by the fact that our method is trained to prevent reliable reconstruction. Moreover, the privacy-sensitive attributes are of lower entropy than the original data, and the compressed representation still contains information leaked about the sensitive attributes. Nevertheless, the results validate that the adversary cannot reliably discover the age attribute based on the protected representation. However, it comes instead as a byproduct but not as a result of design. We should emphasize that preserving of attributes' privacy is not considered in this paper and will be addressed in future work.

IV. Conclusion

In this paper, we considered a problem of privacy-preserving data compression under the adversarial reconstruction and attribute inference attacks. The addressed compression problem is considered as an instance of the variational IB problem in the secure setting. The considered mechanism is based on the ambiguation of the compressed latent space. Using the information advantage of the defender, we formulate the defender's optimization problem under the known secret key. The adversary was assumed to operate under the unknown key trying to train a decoder targeting either adversarial reconstruction of compressed data or adversarial classification of privacy-sensitive attributes. More particularly, we demonstrate that the adversary cannot succeed with the adversarial reconstruction under the properly chosen model parameters. At the same time, the considered privacy-sensitive attributes are of lower entropy than the original data and have different level of correlation with the latent representation. Therefore, we show two examples to demonstrate the adversarial classification. We show that the adversary cannot reliably classify the gender while the binarized age attributes might be leaked from the compressed protected data.

REFERENCES

[1] Naftali Tishby, Fernando C Pereira, and William Bialek, "The information bottleneck method," 2000.

[2] Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard, "From the information bottleneck to the privacy funnel," in *2014 IEEE Information Theory Workshop (ITW 2014)*. IEEE, 2014, pp. 501–505.

[3] Naftali Tishby and Noga Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*. IEEE, 2015, pp. 1–5.

[4] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[5] DJ Strouse and David J Schwab, "The deterministic information bottleneck," *Neural computation*, vol. 29, no. 6, pp. 1611–1630, 2017.

[6] Matias Vera, Leonardo Rey Vega, and Pablo Piantanida, "Collaborative information bottleneck," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 787–815, 2018.

[7] Artemy Kolchinsky, Brendan D Tracey, and Steven Van Kuyk, "Caveats for information bottleneck in deterministic scenarios," in *International Conference on Learning Representations (ICLR)*, 2019.

[8] Seojin Bang, Pengtao Xie, Heewook Lee, Wei Wu, and Eric Xing, "Explaining a black-box using deep variational information bottleneck approach," *arXiv preprint arXiv:1902.06918*, 2019.

[9] Rana Ali Amjad and Bernhard Claus Geiger, "Learning representations for neural network-based classification using the information bottleneck principle," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[10] Jie Hu, Rongrong Ji, ShengChuan Zhang, Xiaoshuai Sun, Qixiang Ye, Chia-Wen Lin, and Qi Tian, "Information competing process for learning diversified representations," in *Advances in Neural Information Processing Systems*, 2019, pp. 2175–2186.

[11] Tailin Wu, Ian Fischer, Isaac L Chuang, and Max Tegmark, "Learnability for the information bottleneck," *International Conference on Learning Representations (ICLR)*, 2019.

[12] Ian Fischer, "The conditional entropy bottleneck," *arXiv preprint arXiv:2002.05379*, 2020.

[13] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata, "Learning robust representations via multi-view information bottleneck," *International Conference on Learning Representations (ICLR)*, 2020.

[14] Alessandro Achille and Stefano Soatto, "Information dropout: Learning optimal representations through noisy computation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2897–2905, 2018.

[15] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *International Conference on Learning Representations (ICLR)*, 2013.

[16] Reza Shokri and Vitaly Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1310–1321.

[17] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman, "Towards the science of security and privacy in machine learning," 2016.

[18] Fatemehsadat Mirshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh, "Privacy in deep learning: A survey," *arXiv preprint arXiv:2004.12254*, 2020.

[19] Seyed Osia, Ali Taheri, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Hamid Rabiee, "Deep private-feature extraction," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, 02 2018.

[20] Michael Veale, Reuben Binns, and Lilian Edwards, "Algorithms that remember: model inversion attacks and data protection law," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, pp. 20180083, 2018.

[21] Han Zhao, Jianfeng Chi, Yuan Tian, and Geoffrey J. Gordon, "Adversarial task-specific privacy preservation under attribute attack," *ArXiv*, vol. abs/1906.07902, 2019.

[22] Elena Zheleva and Lise Getoor, "To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles," 01 2009, pp. 531–540.

[23] Michal Kosinski, David Stillwell, and Thore Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, 03 2013.

[24] Neil Gong and Bin Liu, "You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors," 06 2016.

[25] Neil Gong and Bin Liu, "Attribute inference attacks in online social networks," *ACM Transactions on Privacy and Security*, vol. 21, pp. 1–30, 01 2018.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.