
ASSESSING DIFFERENTIALLY PRIVATE VARIATIONAL AUTOENCODERS UNDER MEMBERSHIP INFERENCE

Daniel Bernau*, Jonas Robl*
 SAP
 Karlsruhe, Germany
 firstname.lastname@sap.com

Florian Kerschbaum
 University of Waterloo
 Waterloo, Canada
 florian.kerschbaum@uwaterloo.ca

April 19, 2022

ABSTRACT

We present an approach to quantify and compare the privacy-accuracy trade-off for differentially private Variational Autoencoders. Our work complements previous work in two aspects. First, we evaluate the the strong reconstruction MI attack against Variational Autoencoders under differential privacy. Second, we address the data scientist’s challenge of setting privacy parameter ϵ , which steers the differential privacy strength and thus also the privacy-accuracy trade-off. In our experimental study we consider image and time series data, and three local and central differential privacy mechanisms. We find that the privacy-accuracy trade-offs strongly depend on the dataset and model architecture. We do rarely observe favorable privacy-accuracy trade-off for Variational Autoencoders, and identify a case where LDP outperforms CDP.

Keywords Variational Autoencoders · Differential Privacy.

1 Introduction

Generative machine learning models such as Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN) infer rules about the distribution of training data to generate new images, tables or numeric datasets that follow the training data distribution. The decision whether to use GAN or VAE depends on the learning task and dataset. However, similar to machine learning models for classification [4, 10, 26, 31, 38] trained generative models leak information about individual training data records [5, 13, 14]. Anonymization of the training data or a training optimizer with differential privacy (DP) can reduce such leakage by limiting the privacy loss that an individual in the training would encounter when contributing their data [1, 3, 17]. Depending on the privacy parameter ϵ differential privacy has a significant impact on the accuracy of the generative model since the perturbation affects how closely generated samples follow the training data distribution. Balancing privacy and accuracy for differentially private generative models is a challenging task for data scientist since privacy parameter ϵ states an upper bound on the privacy loss. In contrast, quantifying the privacy loss under a concrete attack such as membership inference allows to quantify and compare the accuracy-privacy trade-off between differentially private generative models.

This paper compares the privacy-accuracy trade-off for differentially private VAE. This is motivated by previous work that has identified VAE are more prone to membership inference attacks than GAN [14]. Hence, data scientists may want to particularly consider the use of differential privacy when training VAE. In particular, we formulate an experimental study to validate whether our methodology allows to identify sweet spots w.r.t. the privacy-accuracy trade-off in VAE. We conduct experiments for two datasets covering image and activity data, and for three different local and central differential privacy mechanisms. We make the following contributions:

- Quantifying the privacy-accuracy trade-off under membership inference attacks for differentially private VAE.
- Comparing local and central differential privacy w.r.t. the privacy-accuracy trade-off for image and motion data VAE.

* Authors contributed equally.

This paper is structured as follows. Preliminaries are provided in Section 2. We formulate our approach for quantifying and comparing the privacy-accuracy trade-off for DP VAE in Sections 3. Section 4 introduces reference datasets and learning tasks. Section 5 presents the evaluation and is followed by a discussion in Section 6. We discuss related work in Section 7. Section 8 provides conclusions.

2 Preliminaries

In the following we provide preliminaries on VAE, MI and DP.

2.1 Variational Autoencoders

Generative models are trained to learn the joint probability distribution $p(X, Y)$ of features X and labels Y of a training dataset \mathcal{D}^{train} . We focus on Variational Autoencoders (VAE) [22] as generative model. VAE consist of two neural networks: encoder E and decoder D . During training a record x is given to the encoder which outputs the mean $E_\mu(x)$ and variance $E_\Sigma(x)$ of a Gaussian distribution. A latent variable z is then sampled from the Gaussian distribution $N(E_\mu(x), E_\Sigma(x))$ and fed into the decoder D . After successful training the reconstruction $D(z)$ should be close to x . During training two terms are minimized. First, the *reconstruction error* $\|D(z) - x\|$. Second, the *Kullback-Leibler divergence* $KL(N(E_\mu(x), E_\Sigma(x)) || N(0, 1))$ between the distribution of latent variables z and the unit Gaussian. The KL divergence term prevents the network from only memorizing certain latent variables since the distribution should be similar to the unit Gaussian. Kingma et al. [22] motivate the training objective as a lower bound on the log-likelihood and suggest training E and D for a training objective by using the *reparameterization trick*. Samples $D(z)$ are generated from the VAE by sampling a latent variable $z \sim N(0, 1)$ and passing z through D . Similar to GAN conditional VAE generate samples for a specific label by utilizing a condition c as input to E and D .

2.2 Reconstruction Membership Inference Attack against VAE

Membership inference (MI) attacks against machine learning models aim to identify the membership or non-membership of an individual record w.r.t. the training dataset \mathcal{D}^{train} of a target model. To exploit differences in the generated samples of a trained target model the MI adversary \mathcal{A}_{MI} uses a statistical attack model. Therefore, \mathcal{A}_{MI} computes a similarity or error metric for individual records x . After having calculated such a metric for a set of records \mathcal{A}_{MI} labels the records with the highest similarity, or lowest error, as members and all other records as non-members. For VAE the reconstruction loss quantifies how close a reconstructed training record is to the original training data record. Based on the reconstruction loss Hilprecht et al. formulate the reconstruction MI attack against VAE that outperforms prior work [14]. The reconstruction MI attack assumes that a reconstructed training record will have a smaller reconstruction loss than a reconstructed test record and repeatedly computes the reconstruction $\hat{x} = D(z)$ for a record x by drawing the latent variable z from the record-specific latent distribution $\mathcal{N}(\mathbb{E}_\mu(x), \mathbb{E}_\sigma(x))$. The mean reconstruction distance for $N = 300$ samples is then calculated by Eq. (1). Furthermore, the reconstruction MI attack depends on the availability of a distance measure d . In this work we use the generic Mean Squared Error (MSE) and the image domain specific Structural Similarity Index Measure (SSIM) as distance measures. A record x is likely a training record in case of small mean reconstruction distances for MSE or a similarity close to 1 for SSIM.

$$f_{reconstruction} = -\frac{1}{N} \sum_i^N d(\hat{x} - x) \quad (1)$$

2.3 Differential Privacy

For a dataset \mathcal{D} differential privacy (DP) [6] can either be used centrally to perturb a function $f(\mathcal{D})$ or locally to perturb records $x \in \mathcal{D}$ by perturbation. In central DP (CDP) an aggregation function $f(\cdot)$ is first evaluated and then perturbed by a trusted server. Due to perturbation it is no longer possible for an adversary to confidently determine whether $f(\cdot)$ was evaluated on \mathcal{D} , or some neighboring dataset \mathcal{D}' differing in one record. Privacy is provided to records in \mathcal{D} as their impact on $f(\cdot)$ is limited. Mechanisms \mathcal{M} that follow Definition 1 are used for perturbation of $f(\cdot)$ [7]. CDP holds for all possible differences $\|f(\mathcal{D}) - f(\mathcal{D}')\|_2$ by scaling noise to the global sensitivity of Definition 2. To apply CDP in VAE we use a differentially private version [1] of the Adam [21] stochastic gradient optimizer². We refer to this CDP optimizer as DP-Adam. DP-Adam represents a differentially private training mechanism \mathcal{M}_{nn} that updates the weight coefficients θ_t of a neural network per training step $t \in T$ with $\theta_t \leftarrow \theta_{t-1} - \alpha(\tilde{g})$, where $\tilde{g} = \mathcal{M}_{nn}(\partial loss / \partial \theta_{t-1})$ denotes a Gaussian perturbed gradient and α is some scaling function on \tilde{g} to compute an update, i.e., learning rate or

²We used Tensorflow Privacy: <https://github.com/tensorflow/privacy>

running moment estimations. Differentially private noise is added by the Gaussian mechanism of Definition 3. After T update steps, \mathcal{M}_{nn} outputs a differentially private weight matrix θ which is used by the prediction function $h(\cdot)$ of a neural network. DP-Adam bounds the sensitivity of the computed gradients by specifying a clipping norm \mathcal{C} based on which the gradients get clipped before perturbation. The iterative weight updates during training result in a composition of \mathcal{M}_{nn} until training step T at which the final private weights θ are obtained. We measure the privacy loss under composition by composing the Gaussian mechanism σ under Renyi DP [25]. We choose this composition theorem over other composition schemes [1, 19] due to the tighter bound for heterogeneous mechanism invocations. Similar to related work we set $\delta = \frac{1}{|\mathcal{D}|}$ in our experiments [1, 3].

Definition 1 ((ϵ, δ) -Central Differential Privacy). *A mechanism \mathcal{M} gives (ϵ, δ) -central differential privacy if $\mathcal{D}, \mathcal{D}' \subseteq \mathcal{DOM}$ differing in at most one element, and all outputs $\mathcal{S} \subseteq \mathcal{R}$*

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq e^\epsilon \cdot \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta$$

Definition 2 (Global ℓ_2 Sensitivity). *Let \mathcal{D} and \mathcal{D}' be neighboring. The global ℓ_2 sensitivity of a function f , denoted by Δf , is defined as*

$$\Delta f = \max_{\mathcal{D}, \mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_2.$$

Definition 3 (Gaussian Mechanism [8]). *Let $\epsilon \in (0, 1)$ be arbitrary. For $c^2 > 2\ln(\frac{1.25}{\delta})$, the Gaussian mechanism with parameter $\sigma \geq c \frac{\Delta f}{\epsilon}$ gives (ϵ, δ) -CDP, adding noise scaled to $\mathcal{N}(0, \sigma^2)$.*

We refer to the perturbation of records $x \in \mathcal{D}$ as local DP (LDP) [36]. LDP is the standard choice when the server which evaluates a function $f(\mathcal{D})$ is untrusted. In the experiments within this work we use a local randomizer \mathcal{LR} to perturb each record $x \in \mathcal{D}$ independently. Since a record may contain multiple correlated features a \mathcal{LR} must be applied sequentially to each feature which results in a linearly increasing privacy loss. We adapt the definitions of Kasiviswanathan et al. [20] in Definition 4 to achieve LDP by using \mathcal{LR} . A series of \mathcal{LR} executions per record composes to a local algorithm according to Definition 5. ϵ -local algorithms are ϵ -local differentially private [20], where ϵ is a summation of all composed \mathcal{LR} privacy losses. In this work we will use the \mathcal{LR} by Fan [9] for LDP image pixelization. Their \mathcal{LR} applies the Laplace mechanism of Definition 6 with scale $\lambda = \frac{255 \cdot m}{b^2 \cdot \epsilon}$ to each pixel. Parameter m represents the neighborhood in which LDP is provided. Full neighborhood for an image dataset would require that any picture can become any other picture. In general, providing DP or LDP within a large neighborhood will require high privacy parameters ϵ values to retain meaningful image structure. Small privacy parameters ϵ will result in random black and white images.

Definition 4 (Local Differential Privacy). *A local randomizer (mechanism) $\mathcal{LR} : \mathcal{DOM} \rightarrow \mathcal{S}$ is ϵ -local differentially private, if $\epsilon \geq 0$ and for all possible inputs $v, v' \in \mathcal{DOM}$ and all possible outcomes $s \in \mathcal{S}$ of \mathcal{LR}*

$$\Pr[\mathcal{LR}(v) = s] \leq e^\epsilon \cdot \Pr[\mathcal{LR}(v') = s]$$

Definition 5 (Local Algorithm). *An algorithm is ϵ -local if it accesses the database \mathcal{D} via \mathcal{LR} with the following restriction: for all $i \in \{1, \dots, |\mathcal{D}|\}$, if $\mathcal{LR}_1(i), \dots, \mathcal{LR}_k(i)$ are the algorithms invocations of \mathcal{LR} on index i , where each \mathcal{LR}_j is an ϵ_j -local randomizer, then $\epsilon_1 + \dots + \epsilon_k \leq \epsilon$.*

Definition 6 (Laplace Mechanism [8]). *Given a numerical query function $f : \mathcal{DOM} \rightarrow \mathbb{R}^k$, the Laplace mechanism with $\lambda = \frac{\Delta f}{\epsilon}$ is an ϵ -differentially private mechanism, adding noise scaled to $\text{Lap}(\lambda, \mu = 0)$.*

We furthermore use a domain independent LDP mechanism specifically for VAE, to which we refer as VAE-LDP. VAE-LDP by Weggenmann et al. [37] allows a data scientist to use VAE as LDP mechanism to perturb data. This is achieved by limiting the encoders mean and adding noise to the encoders standard deviation before sampling the latent code z during training. After training, the resulting VAE is used to perturb records with $\epsilon = \frac{\Delta f \sqrt{2 \log(1.25/\delta)}}{\sigma}$. In this work we limit the resulting mean of the encoder to $[-3, 3]$ by using the tanh activation function. Furthermore, we introduce noise E according to noise bound σ by enforcing a lower bound on the standard deviation of E . We set the standard deviation to $\max(\sigma, v)$.

3 Accuracy and Privacy for Variational Autoencoders

We compare the privacy-accuracy trade-off for differentially private VAE to support a data scientist \mathcal{DS} in choosing privacy parameters ϵ . For this we formulate a framework to quantify privacy and accuracy as well as the privacy-accuracy trade-off for differentially private VAE with local or central differential privacy. The framework is depicted

in Figure 1. The framework first splits a dataset \mathcal{D} into three distinct subsets: training data \mathcal{D}^{train} , validation data \mathcal{D}^{val} and test data \mathcal{D}^{test} . The *target model* VAE is trained on \mathcal{D}^{train} and optimized on \mathcal{D}^{val} . After training, we use the target model to generate a new dataset \mathcal{D}^{gen} with the same distribution as \mathcal{D}^{train} . We use \mathcal{D}^{gen} as input for the *target classifier*, a feed-forward neural network for classification, to quantify the accuracy of the target model by the target classifier accuracy on \mathcal{D}^{test} . Our framework quantifies privacy by means of a MI adversary \mathcal{A}_{MI} performing a MI attack (cf. Section 2.2). The MI attack dataset \mathcal{D}^{atk} for training and evaluating the MI attack model is sampled equally from \mathcal{D}^{train} and \mathcal{D}^{test} . We use the framework to calculate the baseline trade-off, as well as CDP and LDP trade-off. The baseline trade-off is calculated from the baseline target classifier test accuracy and the MI attack without any DP mechanism. For the CDP trade-off the target model is trained with DP-Adam (cf. Section 2.3).

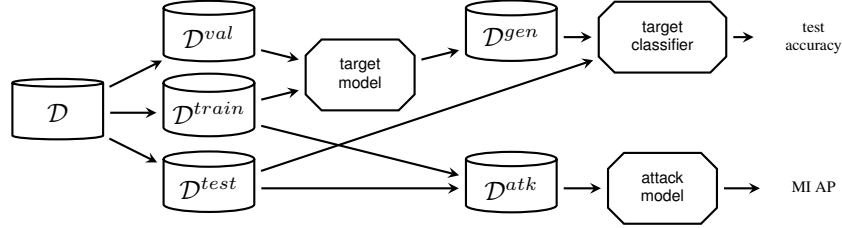


Figure 1: Data flow for the framework.

The LDP trade-off can be computed in three settings to which we refer as LDP-Train, LDP-Full, and VAE-LDP. In LDP-Train a LDP mechanism is applied solely to \mathcal{D}^{train} , but not \mathcal{D}^{val} and \mathcal{D}^{test} . This scheme is similar to Denoising Autoencoders [35]. However, we evaluated the LDP-Train setting and observed it to be mostly impractical for VAE since it introduces a transfer learning task. In particular, working on two different data distributions for \mathcal{D}^{train} and \mathcal{D}^{test} leads to distant latent representations and contrasting reconstructions. This neither benefits the target classifier test accuracy nor reduces MI attack performance in comparison to perturbing both training and test data. Hence, we only mention LDP-Train for sake of completeness but will not discuss LDP-Train in the rest of this work. In LDP-Full, \mathcal{D} is perturbed and the training objective of the target model and the target classifier is changed implicitly (i.e., performance on perturbed data). VAE-LDP perturbs generated data \mathcal{D}^{gen} by training a perturbation model that follows the target model architecture to enforce LDP.

The use of LDP also leads to MI attack variations. In particular, the MI attack can either be evaluated against perturbed or unperturbed records in \mathcal{D}^{atk} . We argue that in the LDP-Full setting the MI attack performance against unperturbed records is particularly relevant from the viewpoint of \mathcal{DS} , since the unperturbed records represent the actual sensitive information and otherwise the attack model would solely learn the to differentiate two distributions by the perturbation skew. Hence, within this work for the LDP settings we exclusively consider the MI attack performance against unperturbed records from \mathcal{D}^{train} .

We evaluate the accuracy of the VAE target model based on the performance of a subsequent target classifier on \mathcal{D}^{test} after training on \mathcal{D}^{gen} . This is a common approach to evaluate the accuracy of generative models [11, 18, 34]. To evaluate the accuracy of the MI attack we use Average Precision of the Precision-Recall curve (MI AP) which considers membership as sensitive information (i.e., neglecting non-membership). The MI AP quantifies the integral under the precision-recall curve as a weighted mean of the precision P per threshold t and the increase in recall R from the previous threshold. Using the accuracy of such a curve instead of a singular value allows us to measure the MI attack performance under optimal conditions. For example, the MI adversary \mathcal{A}_{MI} could decide to increase the assumed certainty by raising the threshold closer to 1. Independently of the target model accuracy, \mathcal{DS} might be interested in lowering MI AP below a predefined threshold that is motivated by legislation (similar to the HIPAA requirement on group sizes [27]).

We quantify the relative trade-off between accuracy and privacy by φ [3] which considers the relative difference between the change in test accuracy for \mathcal{DS} and the change in MI AP for \mathcal{A}_{MI} . We slightly extend the original definition [3] to hold for generic accuracy scores that can be used to quantify the accuracy of the target model as well as success of the attacker. Let ATK be a measure to rate the performance of an attack and ACC a measure to rate the performance of the target model. ATK_{orig} , ACC_{orig} represent the scores without DP, while ATK_{ϵ} , ACC_{ϵ} represent the scores for a specific privacy parameter ϵ . Furthermore, let ATK_{base} , ACC_{base} represent the uniform random guessing baseline where $ATK_{base} = 0.5$ and ACC_{base} depends on the chosen measure. It applies that $ACC_{base} = \frac{1}{C}$ where C depicts the number of classes. Eq. (2) provides our adjusted definition for φ . Similar to the original definition we bound φ between 0 and 2 s.t. φ does not approach infinity when one measure drops while the other remains stable. $0 \leq \varphi \leq 1$ highlights that the relative loss in model accuracy exceeds the relative loss in attack performance. Contrary, for $1 \leq \varphi \leq 2$ the relative loss in model accuracy is smaller than the relative loss in attack performance. In general, a large

gain in privacy, i.e., large drop in attack performance, at a small target model accuracy drop cost is beneficial. Hence \mathcal{DS} seeks to maximize φ .

$$\varphi = \min \left(2, \frac{\max(0, (ATK_{orig} - ATK_{\epsilon}) \cdot (ACC_{orig} - ACC_{base}))}{\max(0, (ACC_{orig} - ACC_{\epsilon}) \cdot (ATK_{orig} - ATK_{base}))} \right) \quad (2)$$

4 Datasets and Learning Tasks

Within this work we use two reference datasets for image and activity data.

Labeled Faces in the Wild (LFW). LFW is a reference dataset for image classification [16]. We resize the 250×250 images to 64×64 by using a bilinear filter and normalize pixels to $[0, 1]$ for improved accuracy. Images are distributed unbalanced across the classes with a minimum of 6 and a maximum of 530 pictures. We consider the most frequent 20 and 50 classes to which we refer as LFW20 and LFW50. In total, LFW20 consists of 1,906 records and LFW50 consists of 2,773 records. 50% of the data is allocated to \mathcal{D}^{train} , 20% to \mathcal{D}^{val} and 30% to \mathcal{D}^{test} . Our VAE target model is an extension of the architecture by Hou et al. [15] and depicted in Figure 2a. E consists of four convolutional layers with 4×4 kernels, a stride of two and Leaky ReLU as activation function. D comprises a dense layer followed by four convolutional layers with 3×3 kernels, a stride of one and Leaky ReLU as activation function. Before each convolutional layer we perform upsampling by a scale of two with the nearest neighbor method. New data is generated by randomly drawing z from a multivariate Gaussian distribution which is passed through the decoder to create a new record. The target classifier is build upon a pre-trained VGG-Very-Deep-16 (VGG16) model [32]. The first part of VGG16 consists of multiple blocks of convolutional layers and max-pooling layers for feature extraction. The second part of VGG16 is a fully-connected network for classification. After loading the pre-trained weights³ we keep the convolutional core and train the classification part.

MotionSense (MS). MS is a reference dataset for human activity recognition with 70610 accelerometer and gyroscope sensor measurements [24]. Each measurement consists of twelve datapoints. Measurements are labeled with activities such as walking downstairs, jogging, and sitting. The associated learning task is to label a time series of measurements collected at 50Hz with the corresponding activity. The VAE target model shall reconstruct such a time series. We normalize the data to $[-1, 1]$ and group the measurements to time series of 10 seconds. 10% of the data is allocated to \mathcal{D}^{train} and \mathcal{D}^{val} each, and the remaining 80% is allocated to \mathcal{D}^{test} . Using 10% of data for training is in line with previous work on MI against generative ML models [5, 13, 14]. For the target model we use a multitask approach in which E consists of a simple LSTM layer with 164 cells followed by two dense layers for μ and σ . μ and σ are used to sample z through the reparameterization trick. D starts with a repeat vector unit for z . This allows us to create sequences and pass z to an LSTM layer. Furthermore, a second LSTM layer with twelve units is used to output sequences for each sensor. To support the reconstruction task we input μ to a classifier. Figure 2b shows the target model architecture. New data is generated by passing training records of a given class E to create z , which is then passed through the decoder to generate a record. We have to sample z from the class-specific latent distribution since the latent space is clustered as a consequence of the multitask classifier. The overall loss is balanced with $\lambda_1 = 0.01$, $\lambda_2 = 50$, $\lambda_3 = 0.5$ for KL-loss, reconstruction loss and classifier loss respectively. The target classifier is based on the Human Activity Recognition Convolutional Neural Network (HARCNN) architecture for time series data by Saeed [29]. In HARCNN each convolutional layer is followed by a dropout layer which we set to 0.3 to learn a more general representation of the data. The final two fully-connected layers are used for classification.

5 Evaluation

Instead of comparing privacy parameter ϵ we designed and performed an experiment to compare the privacy-accuracy trade-off in different DP settings. The experiment quantifies the target classifier test accuracy and MI AP by using the framework depicted in Figure 1 (cf. Section 3). We discuss the experiment for each dataset in four parts. First, we state the *baseline* test accuracy of the target classifier on non-generated data to provide information on the general drop in test accuracy between generated and non-generated data. Second and Third, we discuss CDP and LDP results. Fourth, the results for VAE-LDP are presented. For CDP, LDP and VAE-LDP the experiment results are depicted in two figures each, stating target classifier accuracy over ϵ and MI AP over ϵ . In each figure we also state the original target classifier test accuracy and MI AP for unperturbed data.

³<https://github.com/rcmalli/keras-vggface>

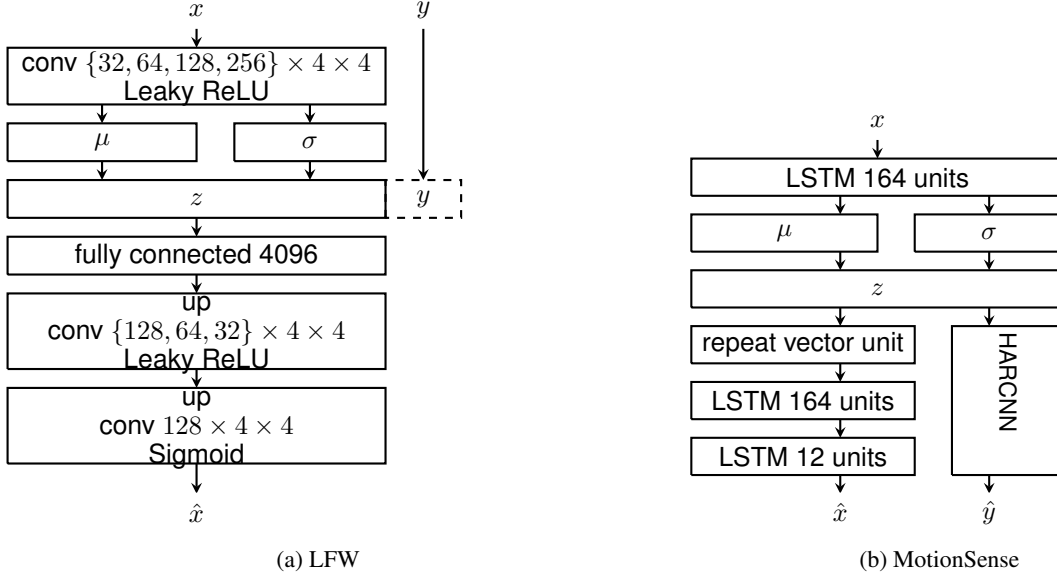


Figure 2: VAE target model architectures.

5.1 Setup

For each dataset the target model is trained for 1000 epochs after which the target model test loss did not decrease significantly while the target classifier accuracy did not increase anymore. The target classifier is trained on generated samples from the VAE until the target classifier test data loss is stagnating (i.e., early stopping). This experiment design avoids overfitting and increases real-world relevance of our results. For CDP we use DP-Adam which samples noise from a Gaussian distribution (cf. Definition 3) with scale $\sigma = \text{noise multiplier } z \times \text{clipping norm } \mathcal{C}$. We use the heuristic of Abadi et al. [1] and set \mathcal{C} as the median of norms of the unclipped gradients over the course of 100 training epochs. We evaluate increasing CDP noise regimes for the target model by evaluating noise multipliers $z \in \{0.001, 0.01, 0.1, 0.5, 1\}$. The noise levels cover a wide range from baseline accuracy to naive majority vote. The exact (ϵ, δ) values are presented in Table 2 in the appendix. Due to the varying LDP mechanisms we state the privacy parameter ϵ_i for a single mechanism execution for feature i per dataset in the next sections and summarize in ϵ in Table 2. VAE-LDP perturbation models are trained with various noise bounds $\sigma \in \{0.1, 1, 10, 100, 1000\}$. Again, the corresponding exact (ϵ, δ) values are presented in Table 2. For the MI attack we randomly draw 1000 records both from $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{test}}$ for \mathcal{D}^{atk} . The experiments were run on Amazon Web Services Elastic Compute Cloud instances of type “p2.xlarge”⁴ with 64 GiB RAM. This instance type is optimized for GPU computing. We implemented our experiments in Python 3.8 and use TensorFlow Privacy⁵. We provide all code on GitHub⁶. We identify hyperparameter values for batch size, epochs and learning rate for all target classifiers with Bayesian optimization.

5.2 LFW

On non-generated baseline images the target classifier achieves baseline test accuracies of 0.78 and 0.66 for LFW20 and LFW50. For generated images we provide two accuracy metrics. Namely, the SSIM of the images generated by the target model and the test accuracy of the target classifier. Figure 4a states the accuracy metrics for unperturbed and CDP perturbed VAE. The figure illustrates that the unperturbed VAE does not generate images with close proximity to the baseline images. However, the images still suffice to produce target classifier test accuracies well above majority voting. Shapes of the head, hair, and some facial expressions as well as the background can be observed for reconstructed images in Figure 6 in the appendix. We also use SSIM as a domain specific distance metric for the reconstruction MI attack. Figure 3b illustrates that the reconstruction MI attack yields a perfect MI AP of 1 for unperturbed VAE. This high MI AP is due to the large gap between train and test SSIM.

⁴<https://aws.amazon.com/ec2>

⁵<https://github.com/tensorflow/privacy>

⁶<https://github.com/SAP-samples/security-research-vae-dp-mia>

Figure 4a states CDP test accuracy over ϵ . The steady accuracy decrease is due to the closing target model train-test gap, which we state in Table 2 in the appendix. The resulting regularization also lowers the SSIM of the generated images. A particular sharp drop in SSIM is observable for $z = 0.5$ ($\epsilon \approx 350$). For this datapoint posterior collapse occurs when E produces noisy μ and σ leading to unstable latent codes z which in turn are ignored by D . In consequence, D produces reconstructions independently of z leading to a increased reconstruction loss, while μ and σ become constant and minimize the KL-loss [23]. As a consequence the target classifier resorts to majority vote. The CDP MI AP over ϵ is stated in Figure 3b. The increased regularization caused by CDP is at the same time lowering MI AP. In addition, due to the inherent label imbalance in LFW the VAE reconstruction of loosely populated classes is worse than the reconstruction for classes with more records. Still, the resulting privacy-accuracy trade-off leaves space for compromise. When \mathcal{DS} would for example be willing to accept an MI AP of up to 0.6 this would require setting $z \leq 0.1$ ($\epsilon \approx 10^5$). $z = 0.1$ leads to target classifier test accuracy of 0.31. However, if \mathcal{DS} raise their threshold to 0.75 this would allow for $z = 0.01$ ($\epsilon \approx 10^8$) and a target classifier test accuracy of 0.52.

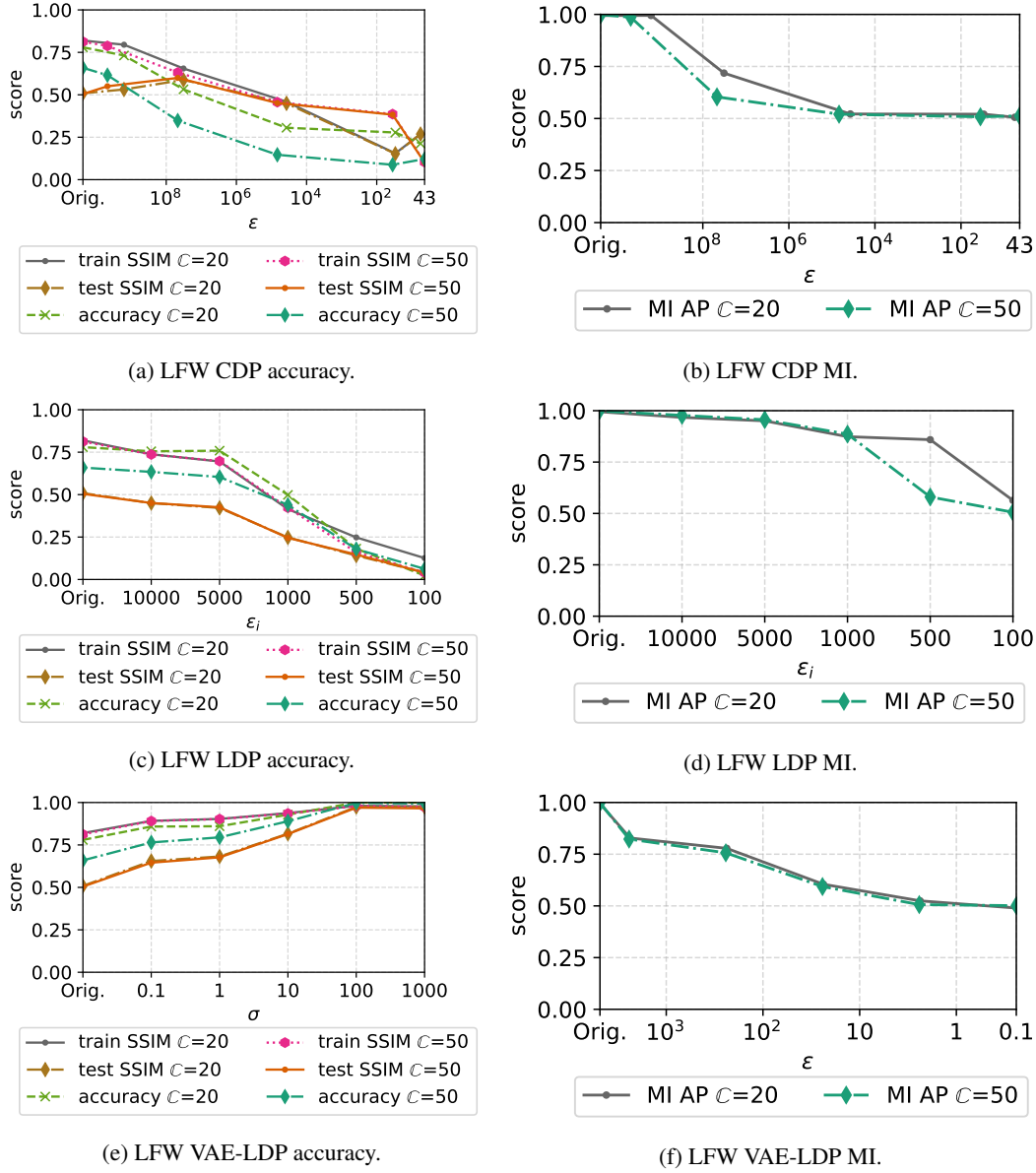


Figure 3: LFW accuracy and privacy.

For LDP we use differentially private image pixelization (cf. Section 2.3) to create LDP training and test datasets with within neighborhood $m = \sqrt{64 \times 64}$. Figure 3c presents the LDP test accuracy and SSIM over ϵ_i . In contrast to the

CDP experiments the target classifier test accuracy and target model SSIM metrics do not show a regularization effect caused by the introduced noise for LDP. The train-test gap narrows only slightly and the random noise introduced in the dataset makes the reconstruction task for the VAE more difficult. Thus, the reconstruction MI attack AP in Figure 3d remains nearly unchanged until $\epsilon_i \leq 500$ at which point the target model SSIM and the target classifier test accuracy are already at poor levels and little room for compromise is existing.

VAE-LDP accuracy over ϵ is presented in Figure 3e. Counterintuitively, the test accuracy even rises over ϵ and the train-test gap and SSIM gap narrow. This is due to the VAE-LDP perturbation model which reconstructs only essential facial features and leaves the background grey when faced with small ϵ . Hence the learning task for the target classifier the reconstruction task for the VAE are simplified. Figure 6 in the appendix underlines this observation by showing the same image for VAE-LDP with increasing noise. The reconstruction attack against VAE-LDP in Figure 3f also decreases as the SSIM gap closes. All in all, the results point towards an advantage of the VAE-LDP mechanism over the LDP image pixelization mechanism. The main disadvantage of the VAE-LDP mechanism over image pixelization is the increased effort to optimize perturbation model hyperparameters.

5.3 MotionSense

Due to the absence of a domain specific accuracy metric we solely consider test accuracy as accuracy metric for this dataset. The target classifier for MS achieves a baseline test-accuracy of 0.99 for non-generated data. Figure 4a states the test accuracy for original and CDP perturbed data over ϵ . The test accuracy is dropping to 0.71 for generated data, which is due to the target model being unable to reconstruct time series for all activities equally well. The reconstruction MI attack has not been used for a time series data in previous work and we suggest to use MSE as reconstruction MI attack distance metric. The original MI attack performance is depicted in Figure 4b and achieves an MI AP 0.52. We see three main reasons for the low MI AP in comparison to LFW. First, MS is more balanced in comparison to LFW. Second, there are significantly more records in MS than in LFW and thus more records per class allow to learn a more general representation. Third, sensor measurements exhibit ambiguities and thus the target model tends to learn more general trends instead of absolute values.

The CDP target classifier test accuracy only slightly worsens with increasing noise as illustrated in Figure 4a. This is mostly due to the target classifier resorting to majority vote for particular activities with increasing noise. Figure 7 in the appendix shows the confusion matrix for the target classifier at $z = 1$ ($\epsilon \approx 16$). The target classifier resorts to majority vote for classes 0 to 3 which represent different types of movements, but is still able to distinguish classes 4 and 5 which represent standing and sitting. The latter two activities are of different nature than the movements and remain distinguishable under noise. The MI AP illustrated in Figure 4b shows again the ineffectiveness of the reconstruction MI attack against the MS time series data.

For LDP we use the Laplace mechanism to perturb each measurement (cf. Section 2.3) and specify the sensitivity per sensor as the maximum of all corresponding observed values to create differentially private time series. Figure 4c shows the target classifier accuracy over ϵ_i . Notably, the target classifier test accuracy increases slightly before dropping sharply over ϵ_i . Here, small noise levels are actually positively influencing the target model training and hence also allow the target classifier to better distinguish between different classes. In general, the simple LDP mechanism used within this experiment seems to prevent the target model to infer structural information and in turn limits reconstruction or and meaningful generation of records. Figure 4d presents the MI attack performance. The MI AP decreases to 0.5 already at the largest ϵ_i and remains close to the baseline for all further ϵ_i .

VAE-LDP test accuracy over ϵ is depicted in Figure 4e. In comparison to LFW the MS perturbation models do not focus on the essential features of the data and in turn the target classifier cannot benefit from increased perturbation. Due to this the predictions also shift to a majority vote for class 5 and lower the test accuracy significantly. The VAE-LDP MI AP over ϵ is illustrated in Figure 4f. Note that at $\sigma = 0.1$ ($\epsilon \approx 40$) an outlier is present where the target model did not learn a continuous latent space and thus the reconstruction of records from \mathcal{D}^{test} suffered. However, the VAE-LDP results show similar trends as the above LDP results.

6 Discussion

This section discusses the findings of this paper w.r.t. comparing the privacy-accuracy trade-off for differentially private VAE.

Image data yields higher MI attack performance than time-series data. The reconstruction MI attack has been shown effective for image data in prior work [5, 14], despite being fairly simple and only taking one metric for disparate behaviour of the target model into consideration. This is in line with the identified gap in image reconstruction for LFW

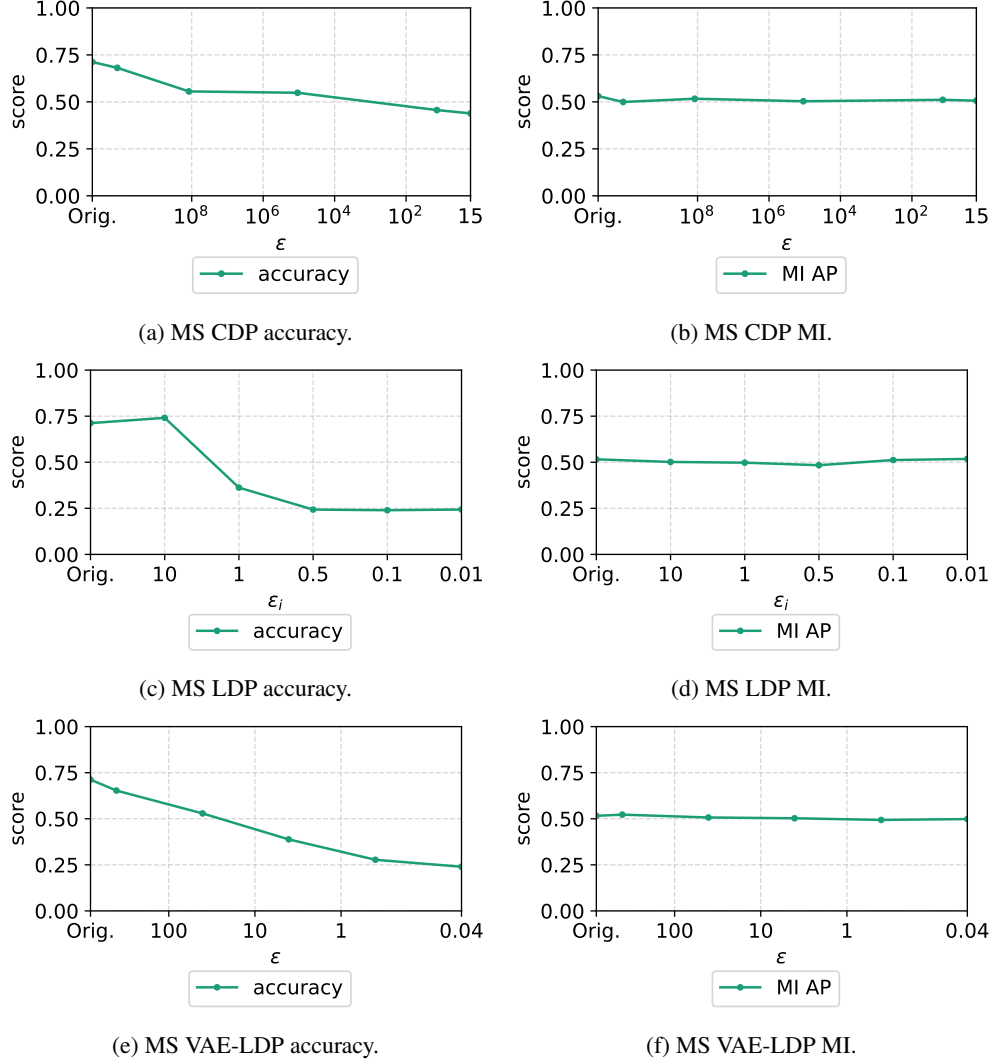


Figure 4: MS accuracy and privacy.

and were able to exploit the gap by using SSIM as a distance measure for reconstruction MI attack. For MS we were not able to identify measure that provides equal success. Since activity measurements exhibit many ambiguities the target model learns to reconstruct relative trends instead of concrete measurements that represent a specific movement. Therefore, the target model generalizes more and is less prone to MI attacks. Additionally, previous research [26, 31] has shown that large datasets with few classes are generally less vulnerable to MI attacks.

Small noise yields favorable relative privacy-accuracy trade-off for image data. For CDP and image data we recommend using as little noise as possible. The relative accuracy drop for \mathcal{DS} largely exceeds the performance loss for \mathcal{A}_M throughout the CDP experiments for LFW. This trend is illustrated in Figure 5a which highlights that the drop in target classifier test accuracy is always larger than the privacy gain by reduced MI AP. For MS the reconstruction MI attack only achieves a performance close to random guessing already against original data. Hence, small DP noise is already sufficient to push the MI AP to random guessing. This is reflected in Figure 5b, where we see an optimal φ already for $z = 0.001$. Similarly for LDP Figures 5c and 5d show only few favorable φ for both datasets and settings. These few favorable trade-offs again indicate that differentially private image pixelization and the Laplace mechanism disproportionately harm model accuracy over protecting privacy. Compared to CDP, LDP shows better trade-offs for small privacy parameter. However, \mathcal{DS} generally gives up more accuracy compared to the gain in privacy.

VAE-LDP outperforms LDP and CDP w.r.t. the relative privacy-accuracy trade-off. In our experiments, the VAE-LDP yielded the best trade-off between target classifier test accuracy and MI AP. This finding is supported by

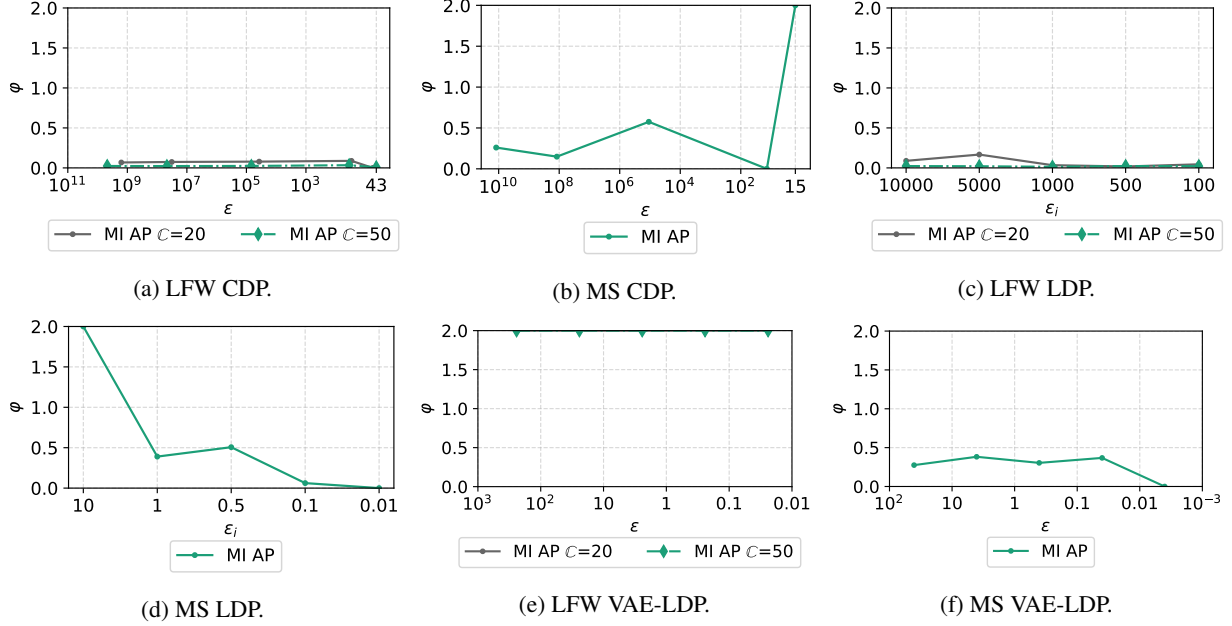


Figure 5: ϕ for CDP, LDP and VAE-LDP, for LFW and MotionSense.

ϕ depicted in Figures 5e and 5f. We identified the interaction between the perturbation models, that retain essential image features, and the targeted classification task as primary reason for the superior trade-off. ϕ for the VAE-LDP experiments highlight that small noise bounds are protecting from the reconstruction MI attack. For larger noise bounds however ϕ only offers limited informative value since the MI AP pivots around random guessing while the target classifier test accuracy is bound by the overall classification baseline.

VAE are highly susceptible to noise introduced during training. Our results indicate that CDP leads to a regularization effect and directly addresses a key driver for MI AP. However, CDP also required additional hyperparameter optimization and increases computational cost. LDP mechanisms consume information within the data to foster protection and hence the test accuracy decrease heavily depends on how the LDP mechanism alters the training data. For example, differentially private image pixelization damages the structures of images to preserve privacy. The more information is consumed by the LDP mechanism, the worse the target classifier test accuracy becomes. This effect is clearly visible for the MS dataset, where the decrease in target classifier accuracy is similar to the overall classification baseline. When this characteristic is present MI is affected mostly as a consequence of diminishing model performance. This is facilitated by the lack of regularization effect which keeps a present relative gap for the MI attacks to exploit. The VAE-LDP mechanism preserves essential features of the LFW dataset during perturbation. The preservation of essential features are beneficial to the overall classification task as the test accuracy remains high while the MI AP decreases.

7 Related Work

We discuss related work from three categories. First, we briefly discuss generative models and accuracy metrics for generative models. Second, we provide background on differential privacy in generative models. Third, we introduce related work on membership inference attacks against generative models.

Generative Adversarial Networks by Goodfellow et al. [12] represent an alternative to VAE. We focus on VAE since VAE in comparison to GAN were observed to be more prone to MI attacks [14]. Salimans et al. [30] introduce Inception Score to automatically evaluate the utility of sampled images from generative models. The main advantage of Inception Score over other metrics such as SSIM is the correlation with human judgements. However, Barrat et al. [2] point out that Inception Score is foremost meaningful for the ImageNet dataset due to pre-training. Therefore, we consider the test accuracy of a target classifier to evaluate the VAE accuracy.

Torkzadehmahani et al. [34] propose the DP-cGAN framework to generate differentially private data and labels. Similar to our work they train target classifiers on the generated data to evaluate model accuracy. We consider VAE with LDP

and CDP. Jordon et al. [18] extend the differentially private federated learning architecture PATE [28] to GAN. Similar to us, they analyze the accuracy of a target classifier for various privacy parameters, yet Jordon et al. do not discuss privacy aside from privacy parameter ϵ . Frigerio et al. [11] evaluate a CDP GAN for time series data also w.r.t. MI attacks. We also consider LDP and quantify the trade-off between privacy and accuracy. Takahashi [33] propose an enhanced version of the DP-SGD for VAE by adjusting the noise that is injected to the loss terms. We use DP-Adam where their improvement is not applicable.

Hayes et al. [13] propose the LOGAN framework for MI attacks against GAN under various assumptions for the knowledge of \mathcal{A}_{MI} . For their black-box attacks they train a separate discriminator model to distinguish between members and non-members. In contrast, we consider statistical MI attack models, allowing for MI attacks against generative models without the need to train a separate attack model. Hilprecht et al. [14] propose Monte-Carlo MI attacks against GAN and VAE. We use their reconstruction MI attack and are the first to consider this attack under differential privacy. Chen et al. [5] extend the reconstruction MI attack to a partial black-box setting where \mathcal{A}_{MI} solely has access to the latent space z but not the internal parameters of the generative model. Their attack composes different losses targeting various aspects of a model and takes the reconstruction as well as the latent representation into consideration. We ran all experiments within this paper also for their attack and the consideration of latent representation did lead to strictly weaker MI AP. The gradient matching attack of Zhu et al. [38] strives for reconstruction of training data from publicly available gradients. In contrast, we focus on the identification of training data.

8 Conclusion

We evaluated a validation framework for quantifying the relative privacy-accuracy trade-off for VAE. We used the framework to compare two LDP and one CDP mechanism for image and time series data w.r.t. their privacy-accuracy trade-off. In particular the LFW image recognition dataset was very susceptible to the reconstruction MI attack whereas the MotionSense activity recognition dataset with more records and less classes was mostly resistant to MI. The CDP mechanism offered a more consistent decrease of MI attack performance whereas the LDP mechanisms showed varying levels of protection depending on chosen privacy parameter and setting. The relative privacy-accuracy trade-off highlights that protection often comes at a disproportionately high accuracy cost.

Appendix

Table 1: Target Classifier hyperparameters.

		Orig., CDP	LDP					VAE-LDP				
			10000	5000	1000	500	100	0.1	1	10	100	1000
LFW20	learning rate	2.4e-05	2.44e-4	8.58e-05	3.66e-05	2.35e-4	1.43e-05	4.03e-4	1.42e-4	9.34e-05	1.39e-4	1.38e-3
	batch size	16	16	16	16	16	64	16	64	64	16	64
	epochs	33	100	10	97	16	24	49	34	50	45	46
	test accuracy	0.98	0.97	0.97	0.82	0.55	0.28	0.94	0.93	0.98	1	1
LFW50	learning rate	1e-05	3.5e-05	4.41e-4	1.85e-4	1.72e-4	1e-05	9.24e-05	3.29e-05	7.39e-05	9.76e-4	1.27e-4
	batch size	16	16	64	64	64	64	16	16	16	64	32
	epochs	100	96	100	35	90	21	49	37	10	32	20
	test accuracy	0.95	0.94	0.93	0.7	0.41	0.2	0.9	0.91	0.97	1	1
MS	ϵ_i		10	1	0.5	0.1	0.01	0.1	1	10	100	1000
	learning rate	9.8e-4	9.37e-4	7.26e-4	7.72e-4	9.87e-05	1.08e-05	1.09e-3	6.75e-4	1.14e-4	2.48e-3	3.71e-05
	batch size	64	64	64	16	16	16	256	128	32	32	64
	epochs	25	25	25	25	25	6	21	9	23	16	24
	test accuracy	0.99	0.98	0.93	0.8	0.29	0.25	0.68	0.53	0.39	0.3	0.24



(a) Reconstructed training records.

(b) Reconstructed test records.



(c) VAE-LDP generated samples for LFW20.

Figure 6: Comparison of reconstructed records and generated samples.

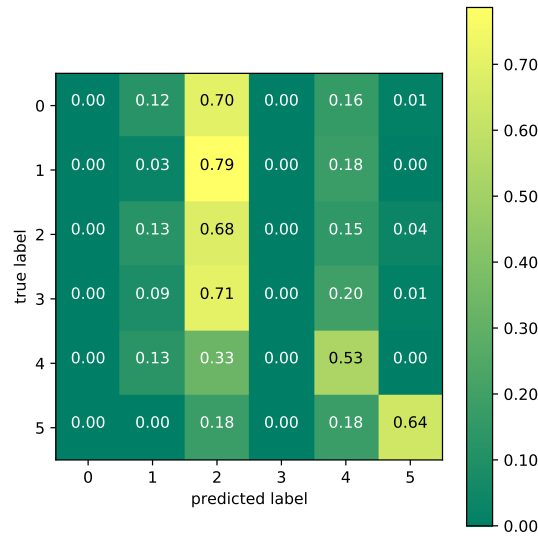


Figure 7: Confusion matrix for the target classifier for MotionSense CDP $z = 1$.

Table 2: Target model hyperparameters, CDP and VAE-LDP (ϵ, δ), and LDP ϵ .

		Orig.	CDP				
z			0.001	0.01	0.1	0.5	1
LFW20	learning rate	5.57e-4	4.67e-4	1.82e-4	1.1e-4	5.62e-4	4.01e-05
	batch size	16	32	16	64	16	32
	epochs	1000	1000				
	\mathcal{C} , microbatch	-	0.03, 4				
	(ϵ, δ)	-	(15948925900.96, 1.08e-03)	(321853746.70, 1.08e-03)	(372950.22, 1.08e-03)	(295.07, 1.08e-03)	(56.41, 1.08e-03)
	train-test gap	260.3	210.6	62.5	13.8	17.9	10.5
LFW50	learning rate	5.88e-4	2.15e-4	1.04e-4	5.14e-05	1.93e-4	6.86e-4
	batch size	16	16		32		
	epochs	1000	1000				
	\mathcal{C} , microbatch	-	0.02, 4				
	(ϵ, δ)	-	(47295786259.73, 7.27e-04)	(468786259.73, 7.27e-04)	(681781.38, 7.27e-04)	(353.74, 7.27e-04)	(43.31, 7.27e-04)
	train-test gap	259.4	195.4	29.4	9.2	7.1	33
MS	learning rate		1e-3				
	batch size		32				
	epochs		1000				
	\mathcal{C} , microbatch	-	3.4e-5, 4				
	(ϵ, δ)	-	(120986947509.93, 1.42e-04)	(1196947509.93, 1.42e-04)	(1093201.38, 1.42e-04)	(137.57, 1.42e-04)	(15.73, 1.42e-04)
	train-test gap	0.7	0.4	0.3	0.1	0.1	0
LDP							
ϵ_i			10000	5000	1000	500	100
LFW20	learning rate		9.22e-4	1.52e-4	2.13e-4	1.14e-4	1e-3
	batch size		32	16	64	32	16
	epochs		1000				
	ϵ		5.718e+07	2.859e+07	5.718e+06	2.859e+06	571800
	train-test gap		267	265	224	160	123
LFW50	learning rate		4.61e-4	2.41e-4	4.31e-4	1.19e-05	1e-05
	batch size		16		64		
	epochs		1000				
	ϵ		8.319e+07	4.1595e+07	8.319e+06	4.1595e+06	831900
	train-test gap		272	264	204	21	5
ϵ_i			10	1	0.5	0.1	0.01
MS	learning rate		1e-3				
	batch size		32				
	epochs		1000				
	ϵ		706190	70619	35309.5	7061.9	706.19
	train-test gap		0.7	0.9	2.7	4.4	4.8
VAE-LDP							
σ			0.1	1	10	100	1000
LFW20	learning rate		5.57e-4				
	batch size		16				
	epochs		1000				
	(ϵ, δ)		(2366.15, 5.25e-04)	(236.61, 5.25e-04)	(23.66, 5.25e-04)	(2.37, 5.25e-04)	(0.24, 5.25e-04)
	train-test gap		156	145	64	3	2
LFW50	learning rate		5.88e-4				
	batch size		16				
	epochs		1000				
	(ϵ, δ)		(2422.52, 3.61e-04)	(242.25, 3.61e-04)	(24.23, 3.61e-04)	(2.42, 3.61e-04)	(0.24, 3.61e-04)
	train-test gap		168	158	68	4	3
MS	learning rate		1e-3				
	batch size		32				
	epochs		1000				
	(ϵ, δ)		(404.96, 1.42e-05)	(40.50, 1.42e-05)	(4.05, 1.42e-05)	(0.40, 1.42e-05)	(0.04, 1.42e-05)
	train-test gap		0.1	0	0.2	0.1	0

References

- [1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep Learning with Differential Privacy. In *Proceedings of Conference on Computer and Communications Security, CCS*. ACM Press, 2016.
- [2] S. Barratt and R. Sharma. A Note on the Inception Score. 2018.
- [3] D. Bernau, J. Robl, P. W. Grassal, S. Schneider, and F. Kerschbaum. Comparing local and central differential privacy using membership inference attacks. In *Proceedings of Conference on Data and Applications Security and Privacy, DBSEC*. Springer, 2021.
- [4] N. Carlini, C. Liu, U. Erlingsson, J. Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *Proceedings of USENIX Security Symposium*. USENIX Association, 2019.
- [5] D. Chen, N. Yu, Y. Zhang, and M. Fritz. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. In *Proceedings of the Conference on Computer and Communications Security, CCS*. ACM Press, 2020.
- [6] C. Dwork. Differential Privacy. In *Proceedings of Colloq. on Automata, Languages and Programming, ICALP*. Springer, 2006.
- [7] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our Data, Ourselves: Privacy via distributed noise generation. In *Proceedings of Conference on Theory and Applications of Cryptographic Techniques (EUROCRYPT)*. Springer, 2006.
- [8] C. Dwork and A. Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), 2014.
- [9] L. Fan. Image pixelization with differential privacy. In *Proceedings of Conference on Data and Applications Security and Privacy, DBSec*. Springer, 2018.
- [10] M. Fredrikson, S. Jha, and T. Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of Conference on Computer and Communications Security, CCS*. ACM Press, 2015.
- [11] L. Frigerio, A. S. de Oliveira, L. Gomez, and P. Duverger. Differentially private generative adversarial networks for time series, continuous, and discrete open data. In *Proceedings of Conference on ICT Systems Security and Privacy Protection*. Springer, 2019.
- [12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Proceedings of Conference on Neural Information Processing Systems, NIPS*. Curran Associates Inc., 2014.
- [13] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro. LOGAN: Membership Inference Attacks Against Generative Models. In *Proceedings on Privacy Enhancing Technologies, PETS*, 2019.
- [14] B. Hilprecht, M. Härterich, and D. Bernau. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. In *Proceedings on Privacy Enhancing Technologies, PETS*, 2019.
- [15] X. Hou, L. Shen, K. Sun, and G. Qiu. Deep feature consistent variational autoencoder. In *Proceedings of Conference on Applications of Computer Vision, (WACV)*. IEEE, 2017.
- [16] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller. Learning to align from scratch. In *Proceedings of Conference on Neural Information Processing Systems, NIPS*. Curran Associates Inc., 2012.
- [17] B. Jayaraman and D. Evans. Evaluating differentially private machine learning in practice. In *Proceedings of USENIX Security Symposium*. USENIX Association, 2019.
- [18] J. Jordon, J. Yoon, and M. V. D. Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *Proceedings of Conference on Learning Representations, ICLR*. IEEE, 2019.
- [19] P. Kairouz, S. Oh, and P. Viswanath. The Composition Theorem for Differential Privacy. *IEEE Transactions on Information Theory*, 63(6), 2017.
- [20] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3), 2008.
- [21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of Conference on Learning Representations, ICLR*. IEEE, 2015.
- [22] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proceedings of Conference on Learning Representations, ICLR*. IEEE, 2014.

- [23] J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi. Dont blame the ELBO! A linear VAE perspective on posterior collapse. In *Proceedings of Conference on Neural Information Processing Systems*, NIPS. Curran Associates Inc., 2019.
- [24] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi. Protecting sensory data against sensitive inferences. In *Proceedings of Workshop on Privacy by Design in Distributed Systems*, W-P2ds. ACM Press, 2018.
- [25] I. Mironov. Rényi differential privacy. In *Proceedings of Computer Security Foundations Symposium (CSF)*. IEEE, 2017.
- [26] M. Nasr, R. Shokri, and A. Houmansadr. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *Proceedings of Symposium on Security and Privacy*, S&P. IEEE, 2019.
- [27] A. D. of Health and H. Services. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule, March 2010.
- [28] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson. Scalable private learning with PATE. In *Proceedings of Conference on Learning Representations*, ICLR. IEEE, 2018.
- [29] A. Saeed. Implementing a CNN for Human Activity Recognition in Tensorflow.
- [30] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Proceedings of Conference on Neural Information Processing Systems*, NIPS. Curran Associates Inc., 2016.
- [31] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *Proceedings of Symposium on Security and Privacy*, S&P. IEEE, 2017.
- [32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of Conference on Learning Representations*, ICLR. IEEE, 2015.
- [33] T. Takahashi, S. Takagi, H. Ono, and T. Komatsu. Differentially Private Variational Autoencoders with Term-wise Gradient Aggregation. 2020.
- [34] R. Torkzadehmahani, P. Kairouz, and B. Paten. DP-CGAN: Differentially Private Synthetic Data and Label Generation. In *Proceedings of Conference on Computer Vision and Pattern Recognition Workshops*, CVPRW. IEEE, 2019.
- [35] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11, 2010.
- [36] T. Wang, J. Blocki, N. Li, and S. Jha. Locally Differentially Private Protocols for Frequency Estimation. In *Proceedings of USENIX Security Symposium*. USENIX Association, 2017.
- [37] B. Weggenmann, V. Rublack, M. Andrejczuk, J. Mattern, and F. Kerschbaum. Dp-vae: Human-readable text anonymization for online reviews with differentially private variational autoencoders. In *Proceedings of Web Conference*, WWW. ACM Press, 2022.
- [38] L. Zhu, Z. Liu, and S. Han. Deep leakage from gradients. In *Proceedings of Conference on Neural Information Processing Systems*, NeurIPS. Curran Associates Inc., 2019.