Federated Generative Privacy

Aleksei Triastcyn , Boi Faltings
Artificial Intelligence Lab
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
{aleksei.triastcyn, boi.faltings}@epfl.ch,

Abstract

In this paper, we propose `FedGP`, a framework for privacy-preserving data release in the federated learning setting. We use generative adversarial networks, generator components of which are trained by `FedAvg` algorithm, to draw privacy-preserving artificial data samples and empirically assess the risk of information disclosure. Our experiments show that `FedGP` is able to generate labelled data of high quality to successfully train and validate supervised models. Finally, we demonstrate that our approach significantly reduces vulnerability of such models to model inversion attacks.

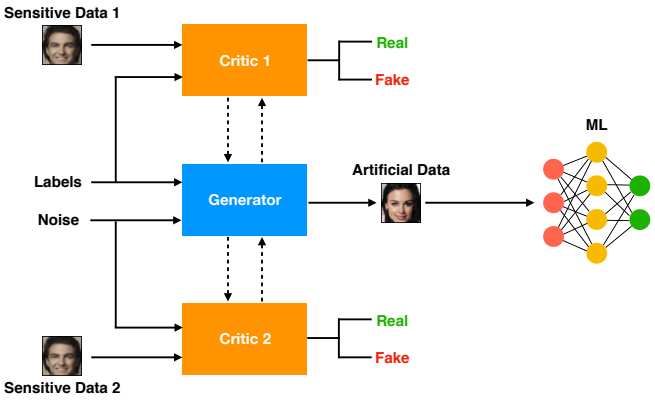

Figure 1: Architecture of our solution for two clients. Sensitive data is used to train a GAN (local critic and federated generator) to produce a private artificial dataset, which can be used by any ML model.

1 Introduction

The rise of data analytics and machine learning (ML) presents countless opportunities for companies, governments and individuals to benefit from the accumulated data. At the same time, their ability to capture fine levels of detail potentially compromises privacy of data providers. Recent research [Fredrikson *et al.*, 2015; Shokri *et al.*, 2017; Hitaj *et al.*, 2017] suggests that even in a black-box setting it is possible to argue about the presence of individual examples in the training set or recover certain features of these examples.

Among methods that tackle privacy issues of machine learning is the recent concept of *federated learning* (FL) [McMahan *et al.*, 2016]. In the FL setting, a central entity (*server*) wants to train a model on user data without actually copying these data from user devices. Instead, users (*clients*) update models locally, and the *server* aggregates these models. One popular approach is the federated averaging, `FedAvg` [McMahan *et al.*, 2016], where *clients* do local on-device gradient descent using their data, then send these updates to the *server* where they get averaged. Privacy can further be enhanced by using secure multi-party computation (MPC) [Yao, 1982] to allow the server access only average updates of a big group of users and not individual ones.

Despite many advantages, federated learning does have a number of challenges. First, the result of FL is a single trained model (therefore, we will refer to it as a *model release* method), which does not provide much flexibility in the future. For instance, it would significantly reduce possibilities for further aggregation from different sources, e.g. different hospitals trying to combine federated models trained on their patients data. Second, this solution requires data to be labelled at the source, which is not always possible, because user may be unqualified to label their data or unwilling to do so. A good example is again a medical application where users are unqualified to diagnose themselves but at the same time would want to keep their condition private. Third, it does not provide provable privacy guarantees, and there is no reason to believe that the aforementioned attacks do not work against it. Some papers propose to augment FL with differential privacy (DP) to alleviate this issue [McMahan *et al.*, 2017; Geyer *et al.*, 2017]. While these approaches perform well in ML tasks and provide theoretical privacy guarantees, they are often restrictive (e.g. many DP methods for ML assume, implicitly or explicitly, access to public data of similar nature or abundant amounts of data, which is not always realistic).

In our work, we address these problems by proposing to combine the strengths of federated learning and recent advancements in generative models to perform privacy-preserving *data release*, which has many immediate advantages. First, the released data could be used to train any ML model (we refer to it as the *downstream task* or the *downstream model*) without additional assumptions. Second, data from different sources could be easily pooled, providing possibilities for hierarchical aggregation and building stronger

models. Third, labelling and verification can be done later down the pipeline, relieving some trust and expertise requirements on users. Fourth, released data could be traded on data markets[1], where anonymisation and protection of sensitive information is one of the biggest obstacles. Finally, data publishing would facilitate transparency and reproducibility of research studies.

The main idea of our approach, named `FedGP`, for *federated generative privacy*, is to train generative adversarial networks (GANs) [Goodfellow *et al.*, 2014] on clients to produce artificial data that can replace clients real data. Since some clients may have insufficient data to train a GAN locally, we instead train a federated GAN model. First of all, user data still remain on their devices. Second, the federated GAN will produce samples from the common cross-user distribution and not from a specific single user, which adds to overall privacy. Third, it allows releasing entire datasets, thereby possessing all the benefits of private *data release* as opposed to *model release*. Figure 1 depicts the schematics of our approach for two clients.

To estimate potential privacy risks, we use our *post hoc* privacy analysis framework [Triastcyn and Faltings, 2019] designed specifically for private data release using GANs.

Our contributions in this paper are the following:

- on the one hand, we extend our approach for private data release to the federated setting, broadening its applicability and enhancing privacy;
- on the other hand, we modify the federated learning protocol to allow a range of benefits mentioned above;
- we demonstrate that downstream models trained on artificial data achieve high learning performance while maintaining good average-case privacy and being resilient to model inversion attacks.

The rest of the paper is structured as follows. In Section 2, we give an overview of related work. Section 3 contains some preliminaries. In Section 4, we describe our approach and privacy estimation framework. Experimental results are presented in Section 5, and Section 6 concludes the paper.

2 Related Work

In recent years, as machine learning applications become a commonplace, a body of work on security of these methods grows at a rapid pace. Several important vulnerabilities and corresponding attacks on ML models have been discovered, raising the need of devising suitable defences. Among the attacks that compromise privacy of training data, model inversion [Fredrikson *et al.*, 2015] and membership inference [Shokri *et al.*, 2017] received high attention.

Model inversion [Fredrikson *et al.*, 2015] is based on observing the output probabilities of the target model for a given class and performing gradient descent on an input reconstruction. Membership inference [Shokri *et al.*, 2017] assumes an attacker with access to similar data, which is used to train a "shadow" model, mimicking the target, and an attack model. The latter predicts if a certain example has already been seen during training based on its output probabilities. Note that both attacks can be performed in a black-box setting, without access to the model internal parameters.

To protect privacy while still benefiting from the use of statistics and ML, many techniques have been developed over the years, including k-anonymity [Sweeney, 2002], l-diversity [Machanavajjhala *et al.*, 2007], t-closeness [Li *et al.*, 2007], and differential privacy (DP) [Dwork, 2006].

Most of the ML-specific literature in the area concentrates on the task of privacy-preserving model release. One take on the problem is to distribute training and use disjoint datasets. For example, [Shokri and Shmatikov, 2015] propose to train a model in a distributed manner by communicating sanitised updates from participants to a central authority. Such a method, however, yields high privacy losses [Abadi *et al.*, 2016; Papernot *et al.*, 2016]. An alternative technique suggested by [Papernot *et al.*, 2016], also uses disjoint training sets and builds an ensemble of independently trained teacher models to transfer knowledge to a student model by labelling public data. This result has been extended in [Papernot *et al.*, 2018] to achieve state-of-the-art image classification results in a private setting (with single-digit DP bounds). A different approach is taken by [Abadi *et al.*, 2016]. They suggest using differentially private stochastic gradient descent (DP-SGD) to train deep learning models in a private manner. This approach achieves high accuracy while maintaining low DP bounds, but may also require pre-training on public data.

A more recent line of research focuses on private data release and providing privacy via generating synthetic data [Bindschaedler *et al.*, 2017; Huang *et al.*, 2017; Beaulieu-Jones *et al.*, 2017]. In this scenario, DP is hard to guarantee, and thus, such models either relax the DP requirements or remain limited to simple data. In [Bindschaedler *et al.*, 2017], authors use a graphical probabilistic model to learn an underlying data distribution and transform real data points (seeds) into synthetic data points, which are then filtered by a privacy test based on a *plausible deniability* criterion. This procedure would be rather expensive for complex data, such as images. Fioretto and Van Hentenryck [2019] employ dicision trees for a hybrid model/data release solution and guarantee stronger ε-differential privacy, but like the previous approach, it would be difficult to adapt to more complex data. Alternatively, Huang *et al.* [2017] introduce the notion of *generative adversarial privacy* and use GANs to obfuscate real data points w.r.t. pre-defined private attributes, enabling privacy for more realistic datasets. Finally, a natural approach to try is training GANs using DP-SGD [Beaulieu-Jones *et al.*, 2017; Xie *et al.*, 2018; Zhang *et al.*, 2018]. However, it proved extremely difficult to stabilise training with the necessary amount of noise, which scales as $\sqrt{m}$ w.r.t. the number of model parameters m. It makes these methods inapplicable to more complex datasets without resorting to unrealistic (at least for some areas) assumptions, like access to public data from the same distribution.

On the other end of spectrum, McMahan *et al.* [2016] proposed federated learning as one possible solution to privacy issues (among other problems, such as scalability and com-

[1] https://www.datamakespossible.com/value-of-data-2018/dawn-of-data-marketplace

munication costs). In this setting, privacy is enforced by keeping data on user devices and only submitting model updates to the server. It can be augmented by MPC [Bonawitz *et al.*, 2017] to prevent the server from accessing individual updates and by DP [McMahan *et al.*, 2017; Geyer *et al.*, 2017] to provide rigorous theoretical guarantees.

3 Preliminaries

This section provides necessary definitions and background. Let us commence with approximate differential privacy.

Definition 1. *A randomised function (mechanism) $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ε, δ)-differential privacy if for any two adjacent inputs $d, d' \in \mathcal{D}$ and for any outcome $o \in \mathcal{R}$ the following holds:*

$$\Pr\left[\mathcal{M}(d) = o\right] \leq e^{\varepsilon} \Pr\left[\mathcal{M}(d') = o\right] + \delta. \tag{1}$$

Definition 2. *Privacy loss of a randomised mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ for inputs $d, d' \in \mathcal{D}$ and outcome $o \in \mathcal{R}$ takes the following form:*

$$L_{(\mathcal{M}(d)\|\mathcal{M}(d'))} = \log \frac{\Pr\left[\mathcal{M}(d) = o\right]}{\Pr\left[\mathcal{M}(d') = o\right]}. \tag{2}$$

Definition 3. *The Gaussian noise mechanism achieving (ε, δ)-DP, for a function $f : \mathcal{D} \to \mathbb{R}^m$, is defined as*

$$\mathcal{M}(d) = f(d) + \mathcal{N}(0, \sigma^2), \tag{3}$$

where $\sigma > C\sqrt{2 \log \frac{1.25}{\delta}}/\varepsilon$ and C is the L2-sensitivity of f.

For more details on differential privacy and the Gaussian mechanism, we refer the reader to [Dwork and Roth, 2014].

In our privacy estimation framework, we also use some classical notions from probability and information theory.

Definition 4. *The Kullback–Leibler (KL) divergence between two continuous probability distributions P and Q with corresponding densities p, q is given by:*

$$D_{KL}(P\|Q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} dx. \tag{4}$$

Note that KL divergence between the distributions of $\mathcal{M}(d)$ and $\mathcal{M}(d')$ is nothing but the expectation of the privacy loss random variable $\mathbb{E}[L_{(\mathcal{M}(d)\|\mathcal{M}(d'))}]$.

Finally, we use the Bayesian perspective on estimating mean from the data to get sharper bounds on expected privacy loss compared to the original work [Triastcyn and Faltings, 2019]. More specifically, we use the following proposition.

Proposition 1. *Let $[l_1, l_2, \ldots, l_n]$ be a random vector drawn from the distribution $p(L)$ with the same mean and variance, and let $\overline{L}$ and S be the sample mean and the sample standard deviation of the random variable L. Then,*

$$\Pr\left(\mathbb{E}[L] > \overline{L} + \frac{F_{n-1}^{-1}(1-\gamma)}{\sqrt{n-1}} S\right) \leq \gamma, \tag{5}$$

where $F_{n-1}^{-1}(1-\gamma)$ is the inverse CDF of the Student's t-distribution with $n-1$ degrees of freedom at $1-\gamma$.

The proof of this proposition can be obtained by using the maximum entropy principle with a flat (uninformative) prior to get the marginal distribution of the sample mean $\overline{L}$, and observing that the random variable $\frac{\mathbb{E}[L]-\overline{L}}{S/\sqrt{n-1}}$ follows the Student's t-distribution with $n-1$ degrees of freedom [Oliphant, 2006].

4 Federated Generative Privacy

In this section, we describe our algorithm, what privacy it can provide and how to evaluate it, and discuss current limitations.

4.1 Method Description

In order to keep participants data private while still maintaining flexibility in downstream tasks, our algorithm produces a federated generative model. This model can output artificial data, not belonging to any real user in particular, but coming from the common cross-user data distribution.

Let $\{u_1, u_2, \ldots, u_n\}$ be a set of *clients* holding private datasets $\{d_1, d_2, \ldots, d_n\}$. Before starting the training protocol, the *server* is providing each *client* with generator G_i^0 and critic C_i^0 models, and *clients* initialise their models randomly. Like in a normal FL setting, the training process afterwords consists of communication rounds. In each round t, *clients* update their respective models performing one or more passes through their data and submit generator updates $\triangle G_i^t$ to the *server* through MPC while keeping C_i^t private. In the beginning of the next round, the *server* provides an updated common generator G^t to all *clients*.

This approach has a number of important advantages:

- Data do not physically leave user devices.
- Only generators (that do not come directly into contact with data) are shared, and critics remain private.
- Using artificial data in downstream tasks adds another layer of protection and limits the information leakage to artificial samples. This is esprecially useful given that ML models can be attacked to extract training data [Fredrikson *et al.*, 2015], sometimes even when protected by DP [Hitaj *et al.*, 2017].

What remains to assess is how much information would an attacker gain about original data. We do so by employing a notion introduced in an earlier work [Triastcyn and Faltings, 2019] that we name *Differential Average-Case Privacy*.

It is important to clarify why we do not use the standard DP to provide stronger theoretical guarantees: we found it extremely difficult to train GANs with the amount of noise required for meaningful DP guarantees. Despite a number of attempts [Beaulieu-Jones *et al.*, 2017; Xie *et al.*, 2018; Zhang *et al.*, 2018], we are not aware of any technically sound solution that would generalise beyond very simple datasets.

4.2 Differential Average-Case Privacy

Our framework builds upon ideas of *empirical DP* (EDP) [Abowd *et al.*, 2013; Schneider and Abowd, 2015] and *on-average KL privacy* [Wang *et al.*, 2016]. The first can be viewed as a measure of sensitivity on posterior distributions

of outcomes [Charest and Hou, 2017] (in our case, generated data distributions), while the second relaxes DP notion to the case of an average user.

More specifically, we say the mechanism $\mathcal{M}$ is (μ, γ)-DAP if for two neighbouring datasets D, D', where data come from an observed distribution, it holds that

$$\Pr(\mathbb{E}[|L_{(\mathcal{M}(D)\|\mathcal{M}(D'))}|] > \mu) \leq \gamma. \quad (6)$$

For the sake of example, let each data point in D, D' represent a single user. Then, $(0.01, 0.001)$-DAP could be interpreted as follows: with probability 0.999, a typical user submitting their data will change outcome probabilities of the private algorithm on average by 1%[2].

4.3 Generative Differential Average-Case Privacy

In the case of generative models, and in particular GANs, we don't have access to exact posterior distributions, a straightforward EDP procedure in our scenario would be the following: *(1)* train GAN on the original dataset D; *(2)* remove a random sample from D; *(3)* re-train GAN on the updated set; *(4)* estimate probabilities of all outcomes and the maximum privacy loss value; *(5)* repeat *(1)–(4)* sufficiently many times to approximate ε, δ.

If the generative model is simple, this procedure can be used without modification. Otherwise, for models like GANs, it becomes prohibitively expensive due to repetitive re-training (steps *(1)–(3)*). Another obstacle is estimating the maximum privacy loss value (step *(4)*). To overcome these two issues, we propose the following.

First, to avoid re-training, we imitate the removal of examples directly on the generated set $\widetilde{D}$. We define a similarity metric $sim(x, y)$ between two data points x and y that reflects important characteristics of data (see Section 5 for details). For every randomly selected real example i, we remove k nearest artificial neighbours to simulate absence of this example in the training set and obtain $\widetilde{D}^{-i}$. Our intuition behind this operation is the following. Removing a real example would result in a lower probability density in the corresponding region of space. If this change is picked up by a GAN, which we assume is properly trained (e.g. there is no mode collapse), the density of this region in the generated examples space should also decrease. The number of neighbours k is defined by the ratio of artificial and real examples, to keep density normalised.

Second, we relax the worst-case privacy loss bound in step *(4)* by the expected-case bound, in the same manner as on-average KL privacy. This relaxation allows us to use a high-dimensional KL divergence estimator [Pérez-Cruz, 2008] to obtain the expected privacy loss for every pair of adjacent datasets ($\widetilde{D}$ and $\widetilde{D}^{-i}$). There are two major advantages of this estimator: it converges almost surely to the true value of KL divergence; and it does not require intermediate density estimates to converge to the true probability measures. Also since this estimator uses nearest neighbours to approximate KL divergence, our heuristic described above is naturally linked to the estimation method.

[2]Because $e^{0.01} \approx 1.01$.

Table 1: Accuracy of student models trained on artificial samples of FedGP compared to non-private centralised baseline and CentGP. In parenthesis we specify the average number of data points per client.

Setting	Dataset	Baseline	CentGP	FedGP
i.i.d.	MNIST (500)	98.10%	97.35%	79.45%
	MNIST (1000)	98.55%	97.39%	93.38%
	MNIST (2000)	98.92%	97.41%	96.23%
non-i.i.d.	MNIST (500)	97.31%	—	83.26%
	MNIST (1000)	98.78%		95.89%
	MNIST (2000)	98.76%		96.88%

Finally, having obtained sufficiently many sample pairs $(\widetilde{D}, \widetilde{D}^{-i})$, we use Proposition 1 to determine DAP parameters μ and γ. This is an improvement over original DAP, because this way we can get much sharper bounds on expected privacy loss.

4.4 Limitations

Our approach has a number of limitations that should be taken into consideration.

First of all, existing limitations of GANs (or generative models in general), such as training instability or mode collapse, will apply to this method. Hence, at the current state of the field, our approach may be difficult to adapt to inputs other than image data. Yet, there is still a number of privacy-sensitive applications, e.g. medical imaging or facial analysis, that could benefit from our technique. And as generative methods progress, new uses will be possible.

Second, since critics remain private and do not leave user devices their performance can be hampered by a small number of training examples. Nevertheless, we observe that even in the setting where some users have smaller datasets overall discriminative ability of all critics is sufficient to train good generators.

Lastly, our empirical privacy guarantee is not as strong as the traditional DP and has certain limitations [Charest and Hou, 2017]. However, due to the lack of DP-achieving training methods for GANs it is still beneficial to have an idea about expected privacy loss rather than not having any guarantee.

5 Evaluation

In this section, we describe the experimental setup and implementation, and evaluate our method on MNIST [LeCun *et al.*, 1998] and CelebA [Liu *et al.*, 2015] datasets.

5.1 Experimental Setting

We evaluate two major aspects of our method. First, we show that training ML models on data created by the common generator achieves high accuracy on MNIST (Section 5.2). Second, we estimate expected privacy loss of the federated GAN and evaluate the effectiveness of artificial data against model inversion attacks on CelebA face attributes (Section 5.3).

Learning performance experiments are set up as follows:

1. Train the federated generative model (*teacher*) on the original data distributed across a number of users.

Table 2: Average-case privacy parameters: expected privacy loss bound μ and probability γ of exceeding it.

Setting	Dataset	μ	γ
i.i.d.	MNIST (500)	0.0117	10^{-15}
	MNIST (1000)	0.0069	
	MNIST (2000)	0.0021	
	CelebA	0.0009	
non-i.i.d.	MNIST (500)	0.0090	10^{-15}
	MNIST (1000)	0.0044	
	MNIST (2000)	0.0020	

2. Generate an artificial dataset by the obtained model and use it to train ML models (*students*).
3. Evaluate students on a held-out test set.

We choose two commonly used image datasets, MNIST and CelebA. MNIST is a handwritten digit recognition dataset consisting of 60000 training examples and 10000 test examples, each example is a 28x28 size greyscale image. CelebA is a facial attributes dataset with 202599 images, each of which we crop to 128x128 and then downscale to 48x48.

In our experiments, we use Python and Pytorch framework.[3] For implementation details of GANs and privacy evaluation, please refer to [Triastcyn and Faltings, 2019]. To train the federated generator we use FedAvg algorithm [McMahan *et al.*, 2016]. As a *sim* function introduced in Section 4.3 we use the distance between InceptionV3 [Szegedy *et al.*, 2016] feature vectors.

5.2 Learning Performance

First, we evaluate the generalisation ability of the student model trained on artificial data. More specifically, we train a student model on generated data and report test classification accuracy on a held-out real set. We compare learning performance with the baseline centralised model trained on original data, as well as the same model trained on artificial samples obtained from the centrally trained GAN (`CentGP`).

Since critics stay private and can only access data of a single user, the size of each individual dataset has significant effect. Therefore, in our experiment we vary sizes of user datasets and observe its influence on training. In each experiment, we specify an average number of points per user, while the actual number is drawn from the uniform distribution with this mean, with some clients getting as few as 100 data points.

We also study two settings: i.i.d. and non-i.i.d data. In the first setting, distribution of classes for each client is identical to the overall distribution. In the second, every client gets samples of 2 random classes, imitating the situation when a single user observes only a part of overall data distribution.

Details of the experiment can be found in Table 1. We observe that training on artificial data from the federated GAN allows to achieve 96.9% accuracy on MNIST with the baseline of 98.8%. We can also see how accuracy grows with the average user dataset size. A less expected observation is that non-i.i.d. setting is actually beneficial for `FedGP`. A possible reason is that training critics with little data becomes easier when this data is less diverse (i.e. the number of different classes is smaller). Comparing to the centralised generative privacy model `CentGP`, we can also see that `FedGP` is more affected by sharding of data on user devices than by overall data size, suggesting that further research in training federated generative models is necessary.

[3] http://pytorch.org

Figure 2: Results of the model inversion attack. Top to bottom: real target images, reconstructions from the non-private model, reconstructions from the model trained by `FedGP`.

Table 3: Face detection and recognition rates (pairs with distances below 0.99) for images recovered by model inversion attack from the non-private baseline and the model trained by `FedGP`.

	Baseline	FedGP
Detection	25.5%	1.2%
Recognition	2.8%	0.1%

5.3 Privacy Analysis

Using the privacy estimation framework (see Sections 4.2 and 4.3), we fix the probability γ of exceeding the expected privacy loss bound μ in all experiments to 10^{-15} and compute the corresponding μ for each dataset and two settings. Table 2 summarises the bounds we obtain. As anticipated, the privacy guarantee improves with the growing number of data points, because the influence of each individual example diminishes. Moreover, the average privacy loss μ, expectedly, is significantly smaller than the typical worst-case DP loss ε in similar settings. To put it in perspective, the average change in outcome probabilities estimated by DAP is $\sim 1\%$ even in more difficult settings, while the state-of-the-art DP method would place the worst-case change at hundreds or even thousands percent without giving much information about a typical case.

On top of estimating expected privacy loss bounds, we test `FedGP`'s resistance to the *model inversion attack* [Fredrikson *et al.*, 2015]. More specifically, we run the attack on two student models: trained on original data samples and on artificial samples correspondingly. Note that we also experimented with another well-known attack on machine learning models, the membership inference [Shokri *et al.*, 2017]. However, we did not include it in the final evaluation, because of the poor attacker's performance in our setting (nearly random guess accuracy for given datasets and models even on the non-private baseline). Moreover, we only consider passive adversaries and we leave evaluation with active adversaries,

e.g. [Hitaj *et al.*, 2017], for future work.

In order to run the attack, we train a student model (a simple multi-layer perceptron with two hidden layers of 1000 and 300 neurons) in two settings: the real data and the artificial data generated by the federated GAN. As facial recognition is a more privacy-sensitive application, and provides a better visualisation of the attack, we pick the CelebA attribute prediction task to run this experiment.

We analyse real and reconstructed image pairs using OpenFace [Amos *et al.*, 2016] (see Table 3). It confirms our theory that artificial samples would shield real data in case of the downstream model attack. In the images reconstructed from a non-private model, faces were detected 25.5% of the time and recognised in 2.8% of cases. For our method, detection succeeded only in 1.2% of faces and the recognition rate was 0.1%, well within the state-of-the-art error margin for face recognition.

Figure 2 shows results of the model inversion attack. The top row presents the real target images. The following rows depict reconstructed images from the non-private model and the model trained on the federated GAN samples. One can observe a clear information loss in reconstructed images going from the non-private to the `FedGP`-trained model. Despite failing to conceal general shapes in training images (i.e. faces), our method seems to achieve a trade-off, hiding most of the specific features, while the non-private model reveals important facial features, such as skin and hair colour, expression, etc. The obtained reconstructions are either very noisy or converge to some average feature-less faces.

6 Conclusions

We study the intersection of federated learning and private data release using GANs. Combined these methods enable important advantages and applications for both fields, such as higher flexibility, reduced trust and expertise requirements on users, hierarchical data pooling, and data trading.

The choice of GANs as a generative model ensures scalability and makes the technique suitable for real-world data with complex structure. In our experiments, we show that student models trained on artificial data can achieve high accuracy on classification tasks. Moreover, models can also be validated on artificial data. Importantly, unlike many prior approaches, our method does not assume access to similar publicly available data.

We estimate and bound the expected privacy loss of an average client by using differential average-case privacy thus enhancing privacy of traditional federated learning. We find that, in most scenarios, the presence or absence of a single data point would not change the outcome probabilities by more than 1% on average. Additionally, we evaluate the provided protection by running the model inversion attack and showing that training with the federated GAN reduces information leakage (e.g. face detection in recovered images drops from 25.5% to 1.2%).

Considering the importance of the privacy research, the lack of good solutions for private data publishing, and the rising popularity of federated learning, there is a lot of potential for future work. In particular, a major direction of advancing current research would be achieving differential privacy guarantees for generative models while still preserving high utility of generated data. A step in another direction would be to improve our empirical privacy concept, e.g. by bounding maximum privacy loss rather than average, or finding a more principled way of sampling from outcome distributions.

References

[Abadi *et al.*, 2016] Martín Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[Abowd *et al.*, 2013] John M Abowd, Matthew J Schneider, and Lars Vilhuber. Differential privacy applications to bayesian and linear mixed model estimation. *Journal of Privacy and Confidentiality*, 5(1):4, 2013.

[Amos *et al.*, 2016] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. 2016.

[Beaulieu-Jones *et al.*, 2017] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *bioRxiv*, page 159756, 2017.

[Bindschaedler *et al.*, 2017] Vincent Bindschaedler, Reza Shokri, and Carl A Gunter. Plausible deniability for privacy-preserving data synthesis. *Proceedings of the VLDB Endowment*, 10(5), 2017.

[Bonawitz *et al.*, 2017] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191. ACM, 2017.

[Charest and Hou, 2017] Anne-Sophie Charest and Yiwei Hou. On the meaning and limits of empirical differential privacy. *Journal of Privacy and Confidentiality*, 7(3):3, 2017.

[Dwork and Roth, 2014] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[Dwork, 2006] Cynthia Dwork. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052, pages 1–12, Venice, Italy, July 2006. Springer Verlag.

[Fioretto and Van Hentenryck, 2019] Ferdinando Fioretto and Pascal Van Hentenryck. Privacy-preserving federated data sharing. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, pages 638–646. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[Fredrikson *et al.*, 2015] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333. ACM, 2015.

[Geyer *et al.*, 2017] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[Hitaj *et al.*, 2017] Briland Hitaj, Giuseppe Ateniese, and Fernando Pérez-Cruz. Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 603–618. ACM, 2017.

[Huang *et al.*, 2017] Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. Context-aware generative adversarial privacy. *Entropy*, 19(12):656, 2017.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Li *et al.*, 2007] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[Machanavajjhala *et al.*, 2007] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[McMahan *et al.*, 2016] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

[McMahan *et al.*, 2017] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

[Oliphant, 2006] Travis E Oliphant. A bayesian perspective on estimating mean, variance, and standard-deviation from data. 2006.

[Papernot *et al.*, 2016] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

[Papernot *et al.*, 2018] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

[Pérez-Cruz, 2008] Fernando Pérez-Cruz. Kullback-leibler divergence estimation of continuous distributions. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 1666–1670. IEEE, 2008.

[Schneider and Abowd, 2015] Matthew J Schneider and John M Abowd. A new method for protecting interrelated time series with bayesian prior distributions and synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(4):963–975, 2015.

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

[Shokri *et al.*, 2017] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 3–18. IEEE, 2017.

[Sweeney, 2002] Latanya Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, October 2002.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[Triastcyn and Faltings, 2019] Aleksei Triastcyn and Boi Faltings. Generating artificial data for private deep learning. In *Proceedings of the PAL: Privacy-Enhancing Artificial Intelligence and Language Technologies, AAAI Spring Symposium Series*, number 2335 in CEUR Workshop Proceedings, pages 33–40, 2019.

[Wang *et al.*, 2016] Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. On-average kl-privacy and its equivalence to generalization for max-entropy mechanisms. In *International Conference on Privacy in Statistical Databases*, pages 121–134. Springer, 2016.

[Xie *et al.*, 2018] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[Yao, 1982] Andrew Chi-Chih Yao. Protocols for secure computations. In *FOCS*, volume 82, pages 160–164, 1982.

[Zhang *et al.*, 2018] Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially private releasing via deep generative model. *arXiv preprint arXiv:1801.01594*, 2018.