Content-Aware Differential Privacy with Conditional Invertible Neural Networks

Malte Tölle[1,3,4(✉)], Ullrich Köthe[2,3], Florian André[1,3,4], Benjamin Meder[1,3,4], and Sandy Engelhardt[1,3,4]

[1] Department of Internal Medicine III, Heidelberg University Hospital, Heidelberg, Germany
malte.toelle@med.uni-heidelberg.de

[2] Visual Learning Lab, Ruprecht-Karls University Heidelberg, Heidelberg, Germany

[3] Informatics for Life Institute, Ruprecht-Karls University Heidelberg, Heidelberg, Germany

[4] DZHK (German Centre for Cardiovascular Research), Partner Site Heidelberg/Mannheim, Heidelberg, Germany

Abstract. Differential privacy (DP) has arisen as the gold standard in protecting an individual's privacy in datasets by adding calibrated noise to each data sample. While the application to categorical data is straightforward, its usability in the context of images has been limited. Contrary to categorical data the meaning of an image is inherent in the spatial correlation of neighboring pixels making the simple application of noise infeasible. Invertible Neural Networks (INN) have shown excellent generative performance while still providing the ability to quantify the exact likelihood. Their principle is based on transforming a complicated distribution into a simple one e.g. an image into a spherical Gaussian. We hypothesize that adding noise to the latent space of an INN can enable differentially private image modification. Manipulation of the latent space leads to a modified image while preserving important details. Further, by conditioning the INN on meta-data provided with the dataset we aim at leaving dimensions important for downstream tasks like classification untouched while altering other parts that potentially contain identifying information. We term our method *content-aware differential privacy* (CADP). We conduct experiments on publicly available benchmarking datasets as well as dedicated medical ones. In addition, we show the generalizability of our method to categorical data. The source code is publicly available at https://github.com/Cardio-AI/CADP.

Keywords: Differential Privacy · Invertible Neural Networks · Normalizing Flows

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-18523-6_9.

S. Albarqouni et al. (Eds.): DeCaF 2022/FAIR 2022, LNCS 13573, pp. 89–99, 2022.
https://doi.org/10.1007/978-3-031-18523-6_9

1 Introduction

The predictive performances of algorithms especially neural networks are heavily dependent on the amount of data they are trained with. In contrast, privacy regulations aiming at hiding individual sensitive information hinder the application of machine learning tools on heterogeneous multi-center data. Since it is not our objective to argue about the benefits of these privacy regulations, we strive to find methods that allow publishing of sensitive data simultaneously to maintaining individual's privacy. While such methods are trivial to implement for categorical data (e.g. a data base with entries for sex, age, gender, etc.) complex data such as images pose a difficult objective. Contrary to categorical data images obtain their meaning by the spatial relationship of individual pixels. Perturbing pixels by adding random noise would not hinder a human or a machine observer from re-identifying the image's content; recognizing people by their face being the most obvious example. Older techniques rely on blurring or pixelation of people's faces, e.g. Google Street View [11].

Training of machine learning models with such samples would tremendously decrease their predictive performance because a great deal of features are lost in the process which the model never sees (see Fig. 1). This is of utmost importance in the medical domain as we must ensure the model learns on valid features for detecting pathologies.

We hypothesize that the tools of machine learning namely neural networks based on Normalizing Flows (NF) known as Invertible Neural Networks (INN) may be used to address the privacy issue when dealing with images and medical ones in particular [2]. Our contribution is three-fold:

- First, we provide mathematically grounded evidence that INNs provide a valuable tool to obtain ϵ-differentially private images that exhibit all features of natural images (e.g. sharpness or authenticity). ϵ quantifies the probability of data leakage, the lower ϵ the more privacy is guaranteed.

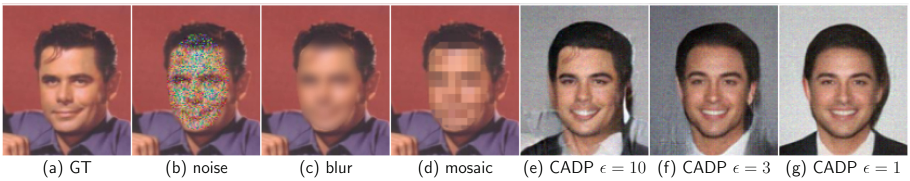


Fig. 1. Example of face anonymization with Differential Privacy [17]. Compared to conventional approaches based on noise (a), blur (b), and mosaic (d) our content-aware approach (e)–(g) changes the identity of the image. For $\epsilon = 10$ (e) one can still see strong similarities between reconstruction and ground truth as e.g. the lock of hair on the forehead. For small ϵ the similarity decreases as desired to disable re-identification. However, if the subsequent task was to classify the eye color, this would still be possible with the CADP results from (e)-(g), since we can condition the transformation and therefore leave important aspects unaltered.

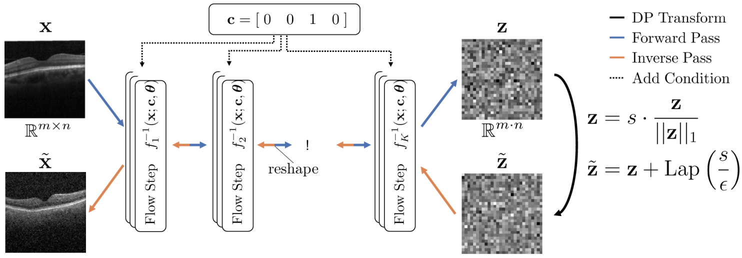
- Second, by conditioning our network on meta-data provided in conjunction with the dataset (e.g. pathologies) the INN is able to automatically extract dimensions most likely corresponding to classifying those meta variables. We assume these features merit attention for downstream tasks and, thus, should be modified as little as possible self-evident within the bounds of desired privacy. We term this method *Content-Aware DP* (CADP).
- Third, we show the generalizability of our method not just to images but also to categorical data making it a universal tool for obtaining differentially private data.

We focus on the task of protecting images in particular, or data in general in any context, detached from their intended usage.

2 Related Work

Differentially Private Invertible Neural Networks. In general each learning based algorithm can be trained in a privacy preserving fashion by using differentially private stochastic gradient descent (DP-SGD) [1]. DP-SGD achieves differentially private model training by clipping the per-sample gradient and adding calibrated Gaussian noise proportional to the desired level of privacy. Therefore, DP-SGD tweaks the model parameters instead of the input to obtain privacy by e.g. ensuring no inputs might be reconstructed from the model parameters [23].

One can distinguish between input-, output-, and algorithm-perturbation to achieve DP. When the output of the algorithm or the algorithm itself is perturbed as e.g. in DP-SGD the analysis is performed on the non-private data, where one has to be concerned about the composition property (ϵ degrades over multiple analyses of the dataset). Further, since one cannot release the data the possibilities for analysis are limited. We circumvent above mentioned limitations by performing input-perturbation and use the robustness of DP against post-processing (any further processing of differential private data retains privacy guarantees).



$\mathbf{x}$
$\mathbb{R}^{m\times n}$
$\tilde{\mathbf{x}}$
$\mathbf{c} = [0 \;\; 0 \;\; 1 \;\; 0]$
Flow Step $f_1^{-1}(\mathbf{x};\mathbf{c},\boldsymbol{\theta})$
Flow Step $f_2^{-1}(\mathbf{x};\mathbf{c},\boldsymbol{\theta})$
!
reshape
Flow Step $f_K^{-1}(\mathbf{x};\mathbf{c},\boldsymbol{\theta})$
$\mathbf{z}$
$\mathbb{R}^{m\cdot n}$
$\tilde{\mathbf{z}}$
DP Transform
Forward Pass
Inverse Pass
Add Condition
$\mathbf{z} = s \cdot \frac{\mathbf{z}}{||\mathbf{z}||_1}$
$\tilde{\mathbf{z}} = \mathbf{z} + \mathrm{Lap}\left(\frac{s}{\epsilon}\right)$

Fig. 2. Content-aware differential privacy (CADP) pipeline. After training the INN to convergence we feed each sample $\mathbf{x}$ with the corresponding condition $\mathbf{c}(\mathbf{y})$ to obtain our latent representation $\mathbf{z}$. After clipping its L_1-norm to the desired sensitivity s, Laplacian distributed noise $\mathrm{Lap}(0, s/\epsilon)$ is added to obtain ϵ-DP. The perturbed $\tilde{\mathbf{z}}$ is fed in reverse to obtain the differentially private image $\tilde{\mathbf{x}}$.

Obviously, INNs can be trained with DP-SGD as well [24]. However, after training one can only use the INN in a generative manner by sampling the latent space $\mathbf{z} \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$ and obtain data samples that have no relation to in reality occuring data samples and are therefore artificial. Thus, it does not allow for perturbation of the real data samples intended to be published or used for model training. Even worse, using artificial data is also not completely secure against attacks [4] and may even lead to wrong pathologies in generated images [5,15].

Differential Privacy for Images. The most prominent application in the literature about differentially private images deals with faces, as this is the most vivid example. Older approaches rely on pixeling, blurring, obfuscation, or inpainting [10], but this has been proven as ineffective against deep learning based recognizers [18,19]. Another promising path is the generation of fully artificial data with e.g. Generative Adversarial Networks (GAN) with the known drawbacks mentioned above [6,21,24,25]. Ziller et al. claimed to having applied DP to medical images. [27]. However, their approach also only involves training a conventional CNN on medical images with DP-SGD. We take a different path and *alter the content of the input image* in a private manner as we want to preserve as much information as possible and only alter dimensions that are not identification related. To the best of our knowledge DP has never been applied *directly to the content of medical images* before.

3 Methods

3.1 (Conditional) Invertible Neural Networks

INNs deal with the approximation of a complex, unobservable distribution $p(\mathbf{x})$ by a simpler tractable prior $q(\mathbf{z})$, usually a spherical multivariate Gaussian. Let $\mathcal{X} = \left\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\right\}$ be n observed i.i.d. samples from $p(\mathbf{x})$. The objective is to approximate $p(\mathbf{x})$ via a model $f_{\boldsymbol{\theta}}$ consisting of a series of K bijective functions $f_{\boldsymbol{\theta}} = f_1 \odot ... \odot f_K$ parameterized fully by $\boldsymbol{\theta}$ transforming $q(\mathbf{z}) = \mathcal{N}(\mathbf{0}; \mathbf{I})$ into $p(\mathbf{x})$ and vice versa ($f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{z} \longleftrightarrow f_{\boldsymbol{\theta}}^{-1}(\mathbf{z}) = \mathbf{x}$).

Such a model can efficiently be used in a generative manner to sample $\mathbf{x} \sim p$ by first sampling $\mathbf{z} \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$ and subsequently transforming the sample as $\mathbf{x} = f_{\boldsymbol{\theta}}(\mathbf{z})$.

Since $f_{\boldsymbol{\theta}}$ exhibits invertibility, exact likelihood evaluation becomes tractable by utilizing the change of variables formula [7,8].

$$\log p(\mathbf{x}) = \log q\left(f_{\boldsymbol{\theta}}^{-1}(\mathbf{z})\right) + \log \left|\det\left(\frac{\partial f_{\boldsymbol{\theta}}^{-1}(\mathbf{z})}{\partial \mathbf{x}}\right)\right| \tag{1}$$

An isotropic Gaussian is usually chosen as prior. Since its covariance matrix is diagonal, components are independent. With INNs sharp image details can be obtained, while simultaneously allowing to modify independent components of the image in latent space [14].

We build on the foundations laid by Ardizzone et al., who incorporated conditions by e.g. concatenation of class labels to the input [3]. This enables the INN to implicitly learn the meta-data dependent distribution in latent space. In the reverse pass we provide the label we would like to obtain, e.g. a pathology, and the INN generates an altered version of the original image that still exhibits the desired pathology ($f_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{c}) = \mathbf{z} \longleftrightarrow f_{\boldsymbol{\theta}}^{-1}(\mathbf{z}, \mathbf{c}) = \mathbf{x}$).

3.2 Content-Aware Differential Privacy

Being termed the gold standard in obscuring data sample sensitive information, DP provides a mathematically grounded, quantifiable measure of leaked information while simultaneously being applicable in a simple manner [26]. From a high-level perspective it guarantees that changing one value in the database ($\mathcal{X}$ and $\mathcal{X}'$) will have only a small effect on the model prediction [9].

$$Pr\left[\mathcal{M}(\mathcal{X}) \in \mathcal{S}\right] \leq \exp(\epsilon) Pr\left[\mathcal{M}(\mathcal{X}') \in \mathcal{S}\right], \tag{2}$$

where $\mathcal{M}$ denotes a randomized mechanism and $\mathcal{S}$ all sets of outputs. The closer the two probabilities are, the less information is leaked (small ϵ). DP is usually obtained by perturbing data with calibrated noise proportional to the function's f (L_1-norm) sensitivity on dataset $\mathcal{X}$, which is the maximum change in the function's value by changing one data point. To achieve pure ϵ-DP the Laplace mechanism is commonly used.

$$s = \max_{\mathcal{X},\mathcal{X}'} ||f(\mathcal{X}) - f(\mathcal{X}')||_1 \,, \tag{3}$$

$$\mathcal{M}(\mathcal{X}) = f(\mathcal{X}) + \text{Lap}\left(\frac{s}{\epsilon}\right) \,. \tag{4}$$

After training an INN to convergence i.e. $f_{\boldsymbol{\theta}}(\mathcal{X}, \mathcal{C}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, each image and label $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X}$ with corresponding condition $\mathbf{c}_i(\mathbf{y}_i)$ is forwarded through the network (see Fig. 2). The resulting latent space $f_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{c}_i(\mathbf{y}_i)) = \mathbf{z}_i$ is modified in a differentially private manner by sampling from a Laplace distribution with standard deviation determined by the sensitivity s and the desired ϵ. We clip our sensitivity by dividing each $\mathbf{z}_i$ by its L_1-norm (Algorithm 1) [1]. Since $\mathcal{Z}$ is learned to be an isotropic Gaussian each component is independent and can, thus, be modified individually. INNs can trivially be expanded to be trained on categorical data as well, making our method a general technique for applying DP on data.

Theorem 1 (ϵ-Content-Aware-DP Mechanism). *For an image $\mathbf{x} \in \mathcal{X}$ there exists a mechanism $\mathcal{M}_{\text{CA}}$ that maps $\mathbf{x}$ to its differentially private counterpart $\tilde{\mathbf{x}} \in \mathcal{X}$. We say $\mathcal{M}_{\text{CA}}$ satisfies ϵ-DP, if and only if for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$*

$$\mathcal{M}_{\text{CA}} = f_{\boldsymbol{\theta}}^{-1}\left[f_{\boldsymbol{\theta}}(\mathbf{x}) + (l_1, ..., l_k)\right] = f_{\boldsymbol{\theta}}^{-1}\left[\mathbf{z} + (l_1, ..., l_k)\right] = f_{\boldsymbol{\theta}}^{-1}\left[\tilde{\mathbf{z}}\right] \,, \tag{5}$$

where $f_{\boldsymbol{\theta}}$ denotes a function that maps $\mathbf{x}$ to a latent vector $\mathbf{z} \in \mathcal{Z}$ and by reverse pass $f_{\boldsymbol{\theta}}^{-1}$ maps $\mathbf{z}$ to $\mathbf{x}$. $\tilde{\mathbf{z}} = \mathbf{z} + (l_1, ..., l_k)$ denotes the ϵ-DP perturbed version of $\mathbf{z}$ with l_i i.i.d. random variables drawn from $\text{Lap}\,(s/\epsilon)$.

Proof. Let $\mathbf{x} \in \mathcal{R}^{|\mathcal{X}|}$ and $\mathbf{x}' \in \mathcal{R}^{|\mathcal{X}|}$ be such that $||\mathbf{x} - \mathbf{x}'||_1 \leq 1$, and $g(\mathbf{x}) = f_{\boldsymbol{\theta}}^{-1}\left(f_{\boldsymbol{\theta}}(\mathbf{x})\right)$ be some function $g : \mathcal{R}^{|\mathcal{X}|} \rightarrow \mathcal{R}^{|\mathcal{Z}|} \rightarrow \mathcal{R}^{|\mathcal{X}|}$. We only consider functions that are volume preserving meaning their Jacobian determinant is equal to one ($|\det\left(\partial f_{\boldsymbol{\theta}}(\mathbf{x})/\partial \mathbf{z}\right)| = 1$). Let $p_{\mathbf{x}}$ denote the probability density function of $\mathcal{M}_{\mathrm{CA}}(\mathbf{x}, g, \epsilon)$, and $p_{\mathbf{x}'}$ of $\mathcal{M}_{\mathrm{CA}}(\mathbf{x}', g, \epsilon)$. We assume the distance between points is similar in $\mathcal{X}$ and $\mathcal{Z}$ as shown by [14]. We compare the two at some arbitrary point $\mathbf{t} \in \mathcal{R}^{|\mathcal{Z}|}$

$$\begin{aligned}
\frac{p_{\mathbf{x}}(\mathbf{t})}{p_{\mathbf{x}'}(\mathbf{t})} &= \prod_{i=1}^{k} \left(\frac{\exp\left(-\frac{\epsilon}{s}|g(\mathbf{x}) - f_{\boldsymbol{\theta}}^{-1}(\mathbf{t})|\right)}{\exp\left(-\frac{\epsilon}{s}|g(\mathbf{x}') - f_{\boldsymbol{\theta}}^{-1}(\mathbf{t})|\right)} \right) = \prod_{i=1}^{k} \left(\frac{\exp\left(-\frac{\epsilon}{s}|f_{\boldsymbol{\theta}}^{-1}\left(f_{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{t}\right)|\right)}{\exp\left(-\frac{\epsilon}{s}|f_{\boldsymbol{\theta}}^{-1}\left(f_{\boldsymbol{\theta}}(\mathbf{x}') - \mathbf{t}\right)|\right)} \right) \\
&= \prod_{i=1}^{k} \left(\exp -\frac{\epsilon}{s}|f_{\boldsymbol{\theta}}^{-1}\left(\mathbf{z}_{\mathbf{x}} - \mathbf{t}\right) - f_{\boldsymbol{\theta}}^{-1}\left(\mathbf{z}_{\mathbf{x}'} - \mathbf{t}\right)| \right) \\
&= \prod_{i=1}^{k} \left(\exp -\frac{\epsilon}{s}|f_{\boldsymbol{\theta}}^{-1}\left(\mathbf{z}_{\mathbf{x}} - \mathbf{z}_{\mathbf{x}'}\right)| \right) \\
&\leq \prod_{i=1}^{k} \exp\left(-\frac{\epsilon|\mathbf{z}_{\mathbf{x}} - \mathbf{z}_{\mathbf{x}'}|}{s} \right) = \exp\left(\frac{\epsilon||\mathbf{z}_{\mathbf{x}} - \mathbf{z}_{\mathbf{x}'}||_1}{s} \right) \\
&\leq \exp(\epsilon),
\end{aligned} \tag{6}$$

where the first inequality follows from the triangle inequality, and the last follows from the definition of sensitivity and $||\mathbf{x} - \mathbf{x}'||_1 \leq 1$. $\frac{p_{\mathbf{x}}(\mathbf{t})}{p_{\mathbf{x}'}(\mathbf{t})} \geq \exp(-\epsilon)$ follows by symmetry.

4 Experiments

We apply our approach for content-aware differential privacy to several publicly available datasets to showcase its generalizability. In each case we first train the INN on the training partition and subsequently train a classifier on the differentially private data. Note that our goal is not to reach as high as possible predictive performance but to close the gap between original and differentially private training. To exemplify the principle of content-aware DP we use the MNIST dataset, since the effect of transformations in latent space is obvious [16]. Next, we use two dedicated medical datasets, the first being a collection of retinal optical coherence tomography (OCT) scans with four classes (choroidal neovascularization (CNV), diabetic macular edema (DME), drusen, and healthy) [12] and the second being a series of chest x-ray scans with healthy and pneumonic patients [12], which contain more complicated and indistinct transformations.

Since most works in adding privacy to images deal with the prototype example of identifiability of faces, we also apply our approach to the CelebA Faces dataset (see Fig. 1) [17]. After having investigated our method on image data, we expand it to categorical data i.e. diabetes dataset from `scikit-learn` [20].

For each dataset we train a separete INN with convolutional subnetworks, with depth (number of downsampling operations) dependent on the image resolution. We chose $d = 2$ for MNIST (28×28), $d = 4$ for OCT and chest x-ray

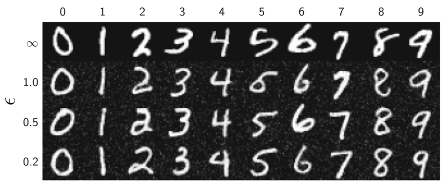


Fig. 3. Differentially private reconstruction of MNIST with different ϵ and $s = \epsilon/2$.

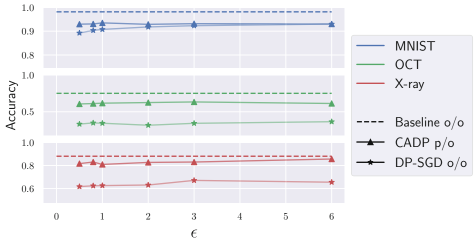


Fig. 4. Accuracy of classifier on different datasets with different ϵ and $s = \min(\epsilon/2, 4)$. Further, we trained the same model with DP-SGD [1]. Training/testing is performed on either original (o) or CADP altered (p) data.

Algorithm 1. CADP

```
Require: Samples from training set $\mathcal{X} = \{(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_N, \mathbf{y}_N)\}$ with corresponding conditions $\mathcal{C} = \{\mathbf{c}_1(\mathbf{y}_1), ..., \mathbf{c}_N(\mathbf{y}_N)\}$, INN $f_\theta$ trained to convergence s.t. $f_\theta(\mathcal{X}) = \mathcal{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, sensitivity $s$, epsilon $\epsilon$
    for $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X}$ and $\mathbf{c}_i(\mathbf{y}_i) \in \mathcal{C}$ do
        Forward pass
        $\mathbf{z}_i \leftarrow f_\theta(\mathbf{x}_i, \mathbf{c}_i(\mathbf{y}_i))$
        Clip norm of $\mathbf{z}_i$
        $\mathbf{z}_i \leftarrow s \cdot \frac{\mathbf{z}_i}{||\mathbf{z}_i||_1}$
        Add calibrated noise
        $\tilde{\mathbf{z}}_i \leftarrow \mathbf{z}_i + \mathrm{Lap}\left(\frac{s}{\epsilon}\right)$
        Reverse Pass
        $\tilde{\mathbf{x}}_i \leftarrow f_\theta^{-1}(\tilde{\mathbf{x}}_i, \mathbf{c}_i(\mathbf{y}_i))$
    end for
Output: $\tilde{\mathcal{X}} = \{(\tilde{\mathbf{x}}_1, \mathbf{y}_1), ..., (\tilde{\mathbf{x}}_N, \mathbf{y}_N)\}$
```

(128 × 128), and $d = 6$ for CelebA (3 × 128 × 128). As coupling block we use the volume preserving GIN (general incompressible-flow) [22] for MNIST and diabetes data, and Glow (generative flow) [14] for the other, more complicated datasets. After having trained an INN to convergence we train a classifier with convolutional blocks and two linear layers on the differentially private data. Testing is performed on original data to investigate the amount of true features the model learns. We believe that the performance of the classifier acts as an implicit benchmark to make sure the INN not only reconstructs conditional noise. It is common practice for all works dealing with DP algorithms to be compared to the non-private benchmark. The goal must be to close the still existing gap to incentivize differentially private training by eliminating all its shortcomings. For comparison we also train the same classifier with DP-SGD, the current gold standard [1]. All experiments were performed on a NVIDIA Titan RTX.

5 Results

The results are presented in a two-fold manner. We first show the differentially private adjusted images per class for each dataset with different levels of ϵ. Second, we show the reached accuracy of the classifier on the original, not-CADP altered test data chunk when trained on the original, on the CADP altered dataset, or with DP-SGD.

MNIST. Even for small ϵ our approach generates visually appealing results that are indistinguishable from real digits but exhibit a large difference from the original (see Fig. 3). Attributes being altered are line thickness (e.g. 6), slant (e.g. 1), and even style (e.g. 2). For $\epsilon = 0.2$ a classifier trained on CADP-altered data outperforms the commonly accepted DP-SGD, CADP reaches 92.94% accuracy while DP-SGD only results in 89.24% (c.f. Fig. 4). The gap closes for larger ϵ.

Retinal OCT and Chest X-ray. In retinal OCTs the perturbations are rather subtle and difficult to interpret for a human observer or a non-expert. Identification related attributes like retinal detachments in specific places are (re-)moved impeding de-identification (see Fig. 5). The CADP-altered images images exhibit transformations resulting in large dissimilarites to their original counterpart. However, CADP induces a smaller privacy-utility tradeoff since the performance of the classifier trained on CADP altered data is close to the one trained on original data (Fig. 4). The classifier trained on data altered by our method outperforms the one trained with DP-SGD by 23.63% on average across all ϵ on the OCT test dataset and by 16.52% on the chest X-ray test dataset. We attribute this to the content-awareness of our method, which leaves dimensions corresponding to conditions, i.e. pathologies, unaltered. This is desirable in settings, where one trains a model on private data of another location, e.g. a hospital, and applies it to its own in-house samples.

Categorical Data. INNs can also generate differentially private categorical data as can be seen in Fig. 6 for the diabetes dataset from `scikit-learn` [20]. The data distributions are kept similar but are still altered equipping each data sample with plausible deniability. To obtain the binary feature of sex, we condition the INN on this feature; the others are learned in an unsupervised fashion.

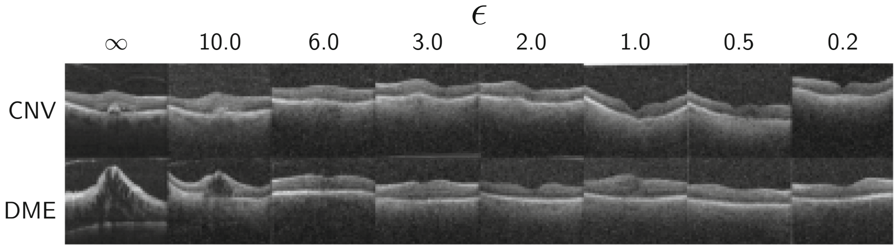


Fig. 5. Content-aware differentially private images from OCT dataset with different ϵ for classes *CNV* and *DME* [12]. The sensitivity is set to min $(\epsilon/2, 4)$. For high ϵ (e.g. 10) the reconstructed retinal OCT still share similarities as in Fig. 1. For smaller ϵ qualitatively the images look different from their original counterpart. However, the classifier (Fig. 4) still performs well acting as an implicit control of the preserved features.

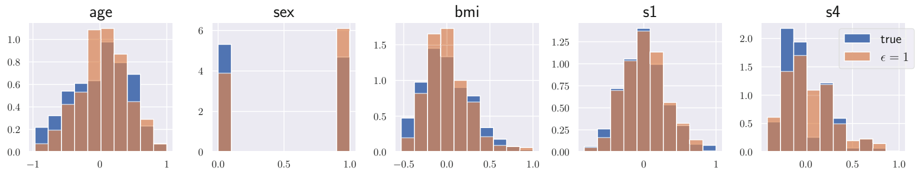


Fig. 6. Content-aware differentially private data from diabetes dataset from `scikit-learn` with $\epsilon = 1$ and sensitivity $s = 1$ [20]. With conditions the INN is able to reconstruct the approximate distributions even if binary distributed.

6 Discussion and Conclusion

We introduced a new method to achieve differentially private images based on invertible neural networks, which we term CADP (content-aware differential privacy). We applied the method to medical images and ensured the identity i.e. pathology of the patient is not changed by conditioning the INN on the class labels. We could show that in three experiments on diverse medical data (images of digits, OCT, and X-ray scans), the subsequent classifiers outperformed conventional approaches by a margin when fed with CADP-generated data. By this we reduce the risk for false diagnosis and increase the safety of patients against wrong diagnoses while providing provable and mathematically grounded privacy guarantees. Hence, CADP pre-processed datasets may be used to increase anonymity of medical image data in the future. However, the level of required anonymity should be decided depending on the individual use case.

Even for small $\epsilon < 1.0$ our method generates visually appealing results that can be used to train a classifier outperforming DP-SGD with the same privacy guarantees. However, clipping of the latent space discards information for reconstruction. In future work, it can be investigated how much information is lost to assure privacy. Further, an in-depth exploration of the latent space can be conducted.

Acknowledgements. This research was supported by grants from the Klaus Tschira Foundation within the Informatics for Life framework, by the DZHK (German Centre for Cardiovascular Research), and by the BMBF (German Ministry of Education and Research). The authors gratefully acknowledge the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG and INST 35/1503-1 FUGG.

References

1. Abadi, M., et al.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (2016). https://doi.org/10.1145/2976749.2978318

2. Ardizzone, L., Kruse, J., Rother, C., Köthe, U.: Analyzing inverse problems with invertible neural networks. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=rJed6j0cKX
3. Ardizzone, L., Lüth, C., Kruse, J., Rother, C., Köthe, U.: Conditional invertible neural networks for guided image generation (2020). https://openreview.net/forum?id=SyxC9TEtPH
4. Bellovin, S., Dutta, P., Reitlinger, N.: Privacy and synthetic datasets. Stan. Technol. Law Rev. (2018)
5. Bhadra, S., Kelkar, V.A., Brooks, F.J., Anastasio, M.A.: On hallucinations in tomographic image reconstruction. IEEE Trans. Med. Imaging **40**, 3249–3260 (2021)
6. Bissoto, A., Perez, F., Valle, E., Avila, S.: Skin lesion synthesis with generative adversarial networks. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pp. 294–302 (2018)
7. Dinh, L., Krueger, D., Bengio, Y.: Nice: non-linear independent components estimation. In: International Conference on Learning Representations (2015)
8. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real NVP. In: International Conference on Learning Representations (2017). https://openreview.net/forum?id=HkpbnH9lx
9. Dwork, C., Roth, A.: Medical imaging deep learning with differential privacy. Sci. Rep. **11**, 1–8 (2021). https://doi.org/10.1038/s41598-021-93030-0
10. Fan, L.: Image pixelization with differential privacy. In: DBSec (2018)
11. Frome, A., et al.: Large-scale privacy protection in google street view. In: International Conference on Computer Vision, pp. 2373–2380 (2009). https://doi.org/10.1109/ICCV.2009.5459413
12. Kermany, D., Zhang, K., Goldbaum, M.: Large dataset of labeled optical coherence tomography (OCT) and chest X-ray images. Cell (2018). https://doi.org/10.17632/rscbjbr9sj.3
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference of Learning Representations (2015)
14. Kingma, D.P., Dhariwal, P.: Glow: generative flow with invertible 1 × 1 convolutions. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
15. Laves, M.H., Tölle, M., Ortmaier, T.: Uncertainty estimation in medical image denoising with Bayesian deep image prior. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis, pp. 81–96 (2020)
16. LeCun, Y., Cortes, C., Burges, C.: MNIST handwritten digit database. ATT Labs, vol. 2 (2010). https://yann.lecun.com/exdb/mnist
17. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: International Conference on Computer Vision (ICCV), December 2015
18. McPherson, R., Shokri, R., Shmatikov, V.: Defeating image obfuscation with deep learning (2016)
19. Oh, S.J., Benenson, R., Fritz, M., Schiele, B.: Faceless person recognition: privacy implications in social media. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 19–35. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_2
20. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
21. Schütte, A.D., et al.: Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. NPJ Digit. Med. **4**, 1–14 (2021). https://doi.org/10.1038/s41746-021-00507-3

22. Sorrenson, P., Rother, C., Köthe, U.: Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN). In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=rygeHgSFDH
23. Usynin, D., et al.: Adversarial interference and its mitigations in privacy-preserving collaborative machine learning. Nat. Mach. Intell. **3**(9), 749–758 (2021). https://doi.org/10.1038/s42256-021-00390-3
24. Waites, C., Cummings, R.: Differentially private normalizing flows for privacy-preserving density estimation. In: AAAI/ACM Conference on AI, Ethics, and Society (2021)
25. Yoon, J., Jordon, J., van der Schaar, M.: PATE-GAN: generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=S1zk9iRqF7
26. Ziller, A., Usynin, D., Braren, R., Makowski, M., Rueckert, D., Kaissis, G.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**, 211–407 (2014). https://doi.org/10.1561/0400000042
27. Ziller, A., Usynin, D., Braren, R., Makowski, M., Rueckert, D., Kaissis, G.: Medical imaging deep learning with differential privacy. Sci. Rep. **11**(1), 1–8 (2021). https://doi.org/10.1038/s41598-021-93030-0