Siamese Generative Adversarial Privatizer for Biometric Data

W. Oleszkiewicz[1], T. Włodarczyk[1], K. Piczak[1], T. Trzcinski[1,3], P. Kairouz[2] and R. Rajagopal[2]
[1]Warsaw University of Technology [2]Stanford University [3]Tooploox
w.oleszkiewicz@ii.pw.edu.pl

Abstract

State-of-the-art machine learning algorithms can be fooled by carefully crafted adversarial examples. As such, adversarial examples present a concrete problem in AI safety. In this work we turn the tables and ask the following question: can we harness the power of adversarial examples to prevent malicious adversaries from learning sensitive information while allowing non-malicious entities to fully benefit from the utility of released datasets? To answer this question, we propose a novel Siamese Generative Adversarial Privatizer that exploits the properties of a Siamese neural network in order to find discriminative features that convey private information. When coupled with a generative adversarial network, our model is able to correctly locate and disguise sensitive information, while minimal distortion constraint prohibits the network from reducing the utility of the resulting dataset. Our method shows promising results on a biometric dataset of fingerprints.

1. Introduction

Large scale datasets enable researchers to design and apply state-of-the-art machine learning algorithms that can solve progressively challenging problems. Unfortunately, most organizations release such datasets rather reluctantly due to the amount of sensitive information they contain about participating individuals.

Ensuring the privacy of subjects is mostly done by anonymizing the dataset before it is made public, *e.g.* by removing personally identifiable information like names or addresses. However, this process is not foolproof. Indeed, correlation and linkage attacks [1, 2] can identify an individual by combining anonymized data with publicly available personal information, *e.g.* in the context of medical data [3].

Typical approaches to countering the shortcomings of anonymization techniques leverage data randomization. While randomizing datasets with differential privacy [4] provides much stronger privacy guarantees, the utility of machine learning models trained on such randomized data is often significantly impaired [5, 6, 7]. We therefore believe that there is an ever increasing need for new privatization methods that preserve the value of the data while protecting the privacy of individuals.

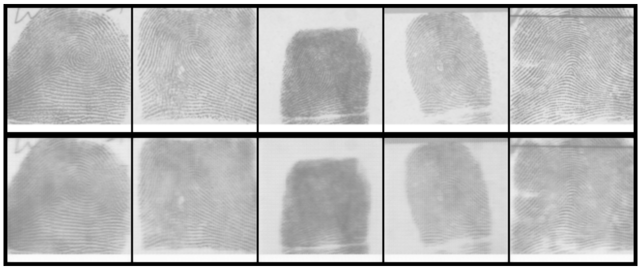

Figure 1. **Top:** Original fingerprints. **Bottom:** Fingerprints with added noise that fool identity discriminator. Our Siamese Generative Adversarial Privatizer learns to locate discriminant image features, such as fingerprint minutiae, and substitutes them.

In this work, we take a new approach towards enabling private data publishing. Instead of adopting worst case, context-free notions of statistical data privacy (such as differential privacy), we present a novel framework that allows the publisher to add noise where it matters. Our framework extends the recent work [8] which presents a Generative Adversarial Privacy (GAP) method that casts the privatization as a constrained minimax game between a privatizer and an adversary that tries to infer private data. The approach we propose here uses a Siamese neural network architecture to identify parts of the data that bear the highest discriminative power and minimally perturb them for privatization. We define empirical conditions that our privatizer needs to fulfill and outline the steps necessary to achieve this goal. Finally, we present promising initial results on a biometric dataset of fingerprint images. Our results show that our privatization framework is able to disguise discriminative characteristics of biometric images with automatically generated adversarial artifacts.

2. Siamase Generative Adversarial Privatizer

The goal of our approach is to develop a privatizer that converts an input image into its randomized version in such a way that: (1) the privacy of the subject is preserved, (2) the utility of the original image is maintained by limiting the amount of distortion added in a context-aware manner, and (3) we can adjust the trade-off between utility and privacy of the resulting data.

To enforce the above conditions, we will use a custom neural network architecture, dubbed *Siamese Generative Adversarial Privatizer*, that comprises two tightly coupled models: a generator $G(\theta_g)$ and a discriminator $D(\theta_d)$. This coupling is inspired by a Siamese architecture [9] and allows the model to identify discriminative parts of data.

For training, the network takes a dataset of n pairs of images $\{I_i, I_i'\}_{i=1}^n$ and their corresponding labels l_i that define whether a pair of images correspond to the same or different identity. In particular, $l_i = 1$ if the images belong to the same person, and $l_i = 0$ if they belong to different persons.

Given a random vector z_i, the generator G tries to create a privatized version of a given image I_i by adding a minimal amount of noise I_{n_i} that can fool the discriminator. On the other hand, the Siamese architecture of the discriminator (verificator) D judges whether a given pair of images belongs to the same subject. This process is represented as a minimax game:

$$\min_{\theta_g} \max_{\theta_d} \frac{1}{n} \sum_{i=1}^{n} l_i \log(D(I_i', I_i + G(z_i;\theta_g));\theta_d)) + \\ (1 - l_i) \log(1 - D(I_i', I_i + G(z_i;\theta_g));\theta_d)). \quad (1)$$

Furthermore, the above minimax optimization problem is subject to the following critical constraint:

$$\frac{1}{n} \sum_{i=1}^{n} d(I_i, I_i + G(z_i;\theta_g)) < \delta, \quad (2)$$

where $d(x, y)$ is a distortion metric and δ is a distortion threshold. The above constraint can be incorporated into the main minimax objective function as follows:

$$\min_{\theta_g} \max_{\theta_d} \sum_{i=1}^{n} l_i \log(D(I_i', I_i + G(z_i;\theta_g));\theta_d)) + \\ (1 - l_i) \log(1 - D(I_i', I_i + G(z_i;\theta_g));\theta_d)) - \\ \lambda \sum_{i=1}^{n} d(I_i, I_i + G(z_i;\theta_g)). \quad (3)$$

We believe that the objective function formulated this way leads the system towards adding minimal amount of noise that is necessary to privatize the data, such that the discriminator is not be able to de-identify the noised image.

3. Experiments

For our experiments, we use NIST 8-Bit Gray Scale Images of Fingerprint Image Groups (FIGS). This database contains 4000 8-bit grayscale fingerprint images paired in couples. Each image is 512-by-512 pixels with 32 rows of white space at the bottom. We use only one image of each pair in our experiments.

Our Siamese Generative Adversarial Privatizer network is trained for 100 epochs using ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use 100-dimensional noise vector as an input to our generator. The discriminator network consists of 3 convolutional layers with kernel size 3, batch normalization and ReLU units, followed by 2 dense layers. We use a dropout factor of 0.2 to regularize the training. The generator's architecture is analogically deconvolutional. The generator and discriminator are trained in tandem. The network is implemented in PyTorch. For our initial experiments, we set the λ parameter in Eq. 3 to 0, hence not imposing any hard constraints on the distortion of the output image.

Fig. 1 shows a sample result obtained as an output of our optimization. The proposed approach locates discriminative fingerprint elements and substitutes them with anonymizing artifacts. By imposing a strict threshold on distortion, we expect to minimize the amount of perturbation and avoid such distinctive artifacts.

As future research, we plan to quantify the trade-off between privacy and utility of the resulting dataset. To that end, we plan to compute the mutual information (MI) between dataset before and after privatization and plot MI values against the results obtained on a proxy task, *e.g.* finger type classification.

References

[1] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 111–125, IEEE, 2008. 1

[2] A. Harmanci and M. Gerstein, "Quantification of private information leakage from phenotype-genotype data: linking attacks," *Nat Meth*, vol. 13, no. 3, pp. 251–256, 2016. 1

[3] L. Sweeney, A. Abu, and J. Winn, "Identifying participants in the personal genome project by name (A re-identification experiment)," *CoRR*, vol. abs/1304.7605, 2013. 1

[4] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*, pp. 1–19, 2008. 1

[5] N. Raval, A. Machanavajjhala, and L. P. Cox, "Protecting visual secrets using adversarial nets," in *CVPRW*, 2017. 1

[6] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Networks," *CoRR*, vol. abs/1705.07663, 2017. 1

[7] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," *CoRR*, vol. abs/1602.07387, 2016. 1

[8] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Context-aware generative adversarial privacy," *CoRR*, vol. abs/1710.09549, 2017. 1

[9] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," in *NIPS*, 1994. 2