Generating Private Data Surrogates for Vision Related Tasks

Ryan Webster, Julien Rabin, Loïc Simon and Frédéric Jurie

Normandie Univ., ENSICAEN, UNICAEN, CNRS, GREYC, France

Motivation & Overview

Data surrogate $\mathcal{D}'$ is obtained by combining image samples from a generator network G and associating them with plausible labels obtained from a classifier C trained on the private train dataset $\mathcal{D}_\mathcal{T}$. A privacy preserving classifier C' is then obtained, displaying similar performance and accuracy on a separate validation set $\mathcal{D}_\mathcal{V}$. The obtained public dataset $\mathcal{D}'$ (and by composition the network C') is robust to membership attack described in Alg 1.

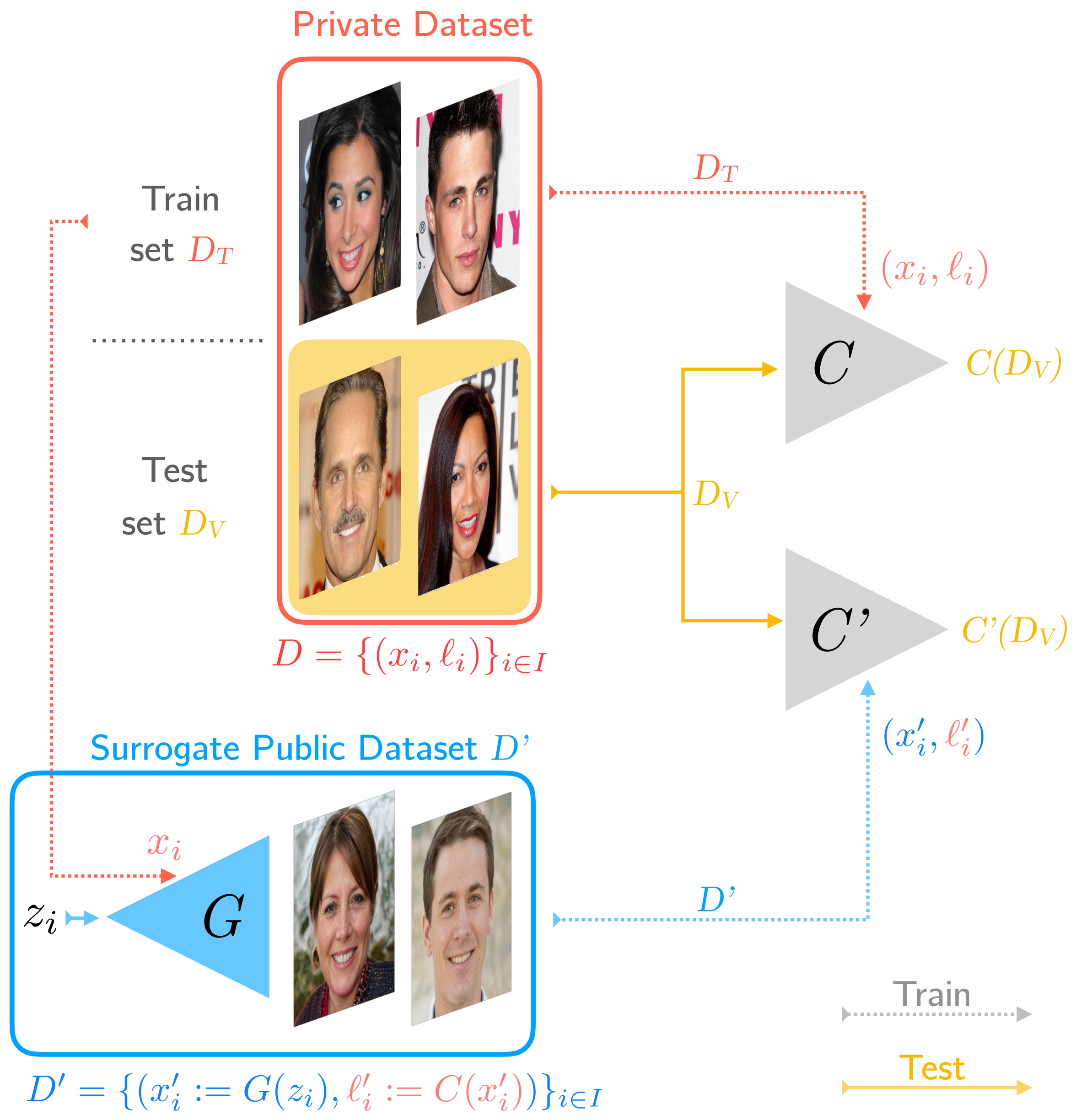


Figure 1: Overview of the proposed framework for creating private data surrogates and its application to train a private task-driven network.

Algorithm 1 Membership attack

Input: Training set $\mathcal{D}_\mathcal{T}$, validation set $\mathcal{D}_\mathcal{V}$

1: Set the attack score function A, either as the recovery loss f_G in Eq. (1) or as the discriminator D.

2: Let $x_i \in \mathcal{D}_\mathcal{T} \cup \mathcal{D}_\mathcal{V}$, such that

$$\begin{cases} x_i \in \mathcal{D}_\mathcal{T} & \text{if } i \leq N \\ x_i \in \mathcal{D}_\mathcal{V} & \text{if } N+1 \leq i \leq 2N \end{cases}$$

3: Sorted indices: $I \leftarrow \operatorname{argsort}\{A(x_i)\}_{1 \leq i \leq 2N}$

Output:

4: Estimated set of training images: $\mathcal{T} \leftarrow \{x_{I(i)}\}_{1 \leq i \leq N}$

5: Membership attack accuracy: $Acc \leftarrow |I \cap \{i : 1 \leq i \leq N\}|/N$

The latent recovery loss for a given image $x_i \in \mathcal{D}_\mathcal{T} \cup \mathcal{D}_\mathcal{V}$ is

$$f_G(x_i) := \|\phi(G(E(x_i))) - \phi(x_i)\|_2^2 \quad (1)$$

where E is an Encoder Network (trained on generated images $G(z_i)$) and ϕ perceptual (e.g. VGG) features.

Evaluation of Performance with Generated Surrogate Datasets

Evaluation: Classification accuracy on **CelebA-HQ** dataset (Table 1) and regression precision on **UTK-Face** dataset (Table 2).

Conclusion: Classifier C' trained on surrogate datasets performs as well as the private one C on the private validation set $\mathcal{D}_\mathcal{V}$.

	CelebA-HQ	Gender	Smiling	Average (5 attributes)	Change in Performance	FID
C	Real Data	94.50	85.20	90.64	-	-
C'	DCGAN	91.90	82.10	86.50	4.14	67.07
	MESCH	92.60	81.45	88.90	1.74	26.31
	LSGAN	92.10	80.80	88.35	2.29	42.01
	PGGAN	93.10	83.05	89.35	**1.29**	**19.17**

Table 1: *Performance of various surrogate datasets on the* **CelebA-HQ** *binary attribute recognition task. Top row represents a classifier C trained on the original dataset $\mathcal{D}_\mathcal{T}$, subsequent rows represent classifiers C' trained with GAN images that are labelled with C. Accuracy represents percent correct on a validation set $\mathcal{D}_\mathcal{V}$. FID scores are reported in the last column (lower is better) to assess the quality of generated images.*

	UTK-Face	Age (MAD error, in years)	Change in Performance (in years)	FID
C	Real Data	5.22	-	-
C'	DCGAN	12.03	6.81	89.68
	LSGAN	5.56	0.34	31.05
	PGGAN	5.12	**-0.10**	**30.65**

Table 2: *Performance of various surrogate datasets on the age regression task of* **UTK-Face**.

Evaluation of Robustness to Membership Attack

Evaluation: Membership attack using Algorithm 1 on **CelebA-HQ** and **UTK-Face** datasets.

Conclusion: Membership attacks are not efficient when a GAN is trained with sufficient data. Membership attacks based on the discriminative network are more efficient, yet a fairly unrealistic scenario.

CelebA-HQ	L_2 Recovery	VGG-Face Recovery	VGG-19 Recovery	Discriminator D
DCGAN	54.1	54.5	51.6	57.1
MESCH	53.9	50.8	52.5	50.1
LSGAN ($\|\mathcal{D}_\mathcal{T}\| = 26k$)	54.8	54.1	54.0	62.9
LSGAN ($\|\mathcal{D}_\mathcal{T}\| = 5k$)	58.1	56.2	57.8	**99.4**
PGGAN	52.0	50.3	52.1	N/A

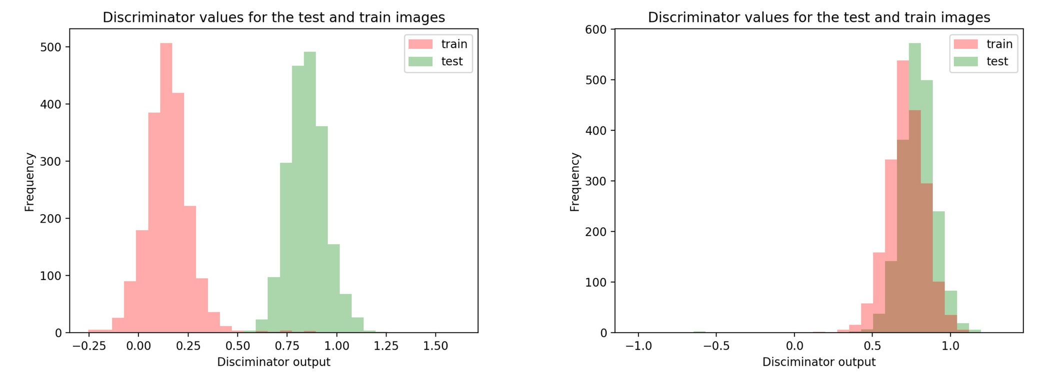
Table 3: *Membership attack accuracies (in %) for various GAN methods trained on the* **CelebA-HQ** *dataset and various attack methods (see Algorithm 1). When not specified otherwise, the size of the training dataset is $|\mathcal{D}_\mathcal{T}| = 26k$ and for the validation set $|\mathcal{D}_\mathcal{V}| = 2k$. GAN methods are reported in the first column. The next three columns use latent recovery attack with loss function f_G (see Eq. 1), with ϕ taken to be the identity, VGG-Face or VGG-19 features respectively. The final column reports the discriminative attack accuracy with the discriminator D from the GAN training (the discriminator of PGGAN requires feeding a whole batch which prevented us to implement this attack). As a baseline, the same discriminative attack is done on LSGAN with a smaller training dataset (5k) demonstrating that in such setting the discriminator network is capable of memorizing almost perfectly the entire training dataset.*

UTK-Face	L_2 Recovery	VGG-Face Recovery	VGG-19 Recovery	Discriminator
DCGAN	52.3	53.5	52.1	50.9
LSGAN	53.4	53.9	53.6	75.8
PGGAN	54.7	56.8	54.1	-

Table 4: *Membership attack accuracies (in %) for various GAN methods trained on the* **UTK-Face** *dataset.*

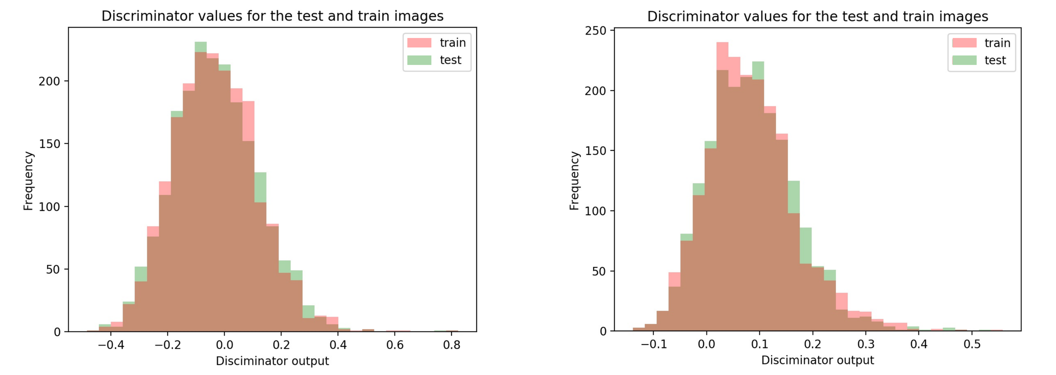
Visual Results

$|\mathcal{D}_\mathcal{T}| = 5,000$ $\quad$ $|\mathcal{D}_\mathcal{T}| = 26,000$



Acc = **99.4**% $\quad$ Acc = 62.9%

$|\mathcal{D}_\mathcal{T}| = 5,000$ $\quad$ $|\mathcal{D}_\mathcal{T}| = 26,000$



Acc = 50% $\quad$ Acc = 50.1%

Figure 2: Histogram of attack scores based on the Discriminator D for $N = 2000$ images from the training set $\mathcal{D}_\mathcal{T}$ (in red) and the test set $\mathcal{D}_\mathcal{V}$ (in green) for LSGAN (first two rows) and MESCH (next two rows) trained on CelebA-HQ, respectively with $|\mathcal{D}_\mathcal{T}| = 5,000$ images (left column) and 26,000 images (right column). While the quality of images does not improve a lot with a larger number of training images, the robustness to discriminative attack increases dramatically for LSGAN (average membership inference attack accuracy are given in the last row).

Acknowledgements

This work was supported by Region Normandie, under grant RIN NormanD'eep.