Attribution of Private Images to GANs using Differential Privacy

Aravind Telidevara, Bhanu Prakash Reddy Chennam, Satabdhi Reddy Nalamalpu, and Sara Ashcraft

University of North Carolina at Charlotte

1. Abstract

The sanitization of sensitive content in image data, e.g., faces and texts, has largely relied on standard image obfuscation techniques, such as pixelization (also referred to as Mosaicing) and blurring. However, recent developments and applications of machine learning have rendered such obfuscation techniques ineffective for privacy preservation. Moreover, Convolutional Neural Networks (CNN) are highly adaptable to standard obfuscation.

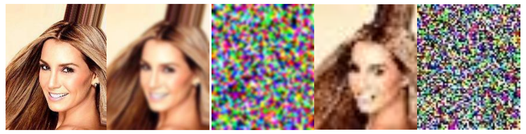
We tried to check the capability of the fingerprint classifier in classifying the proGAN and real images ,when the images are perturbed with various methods like gaussian blur, pixelization and their Differential Privacy versions.

2. Dataset

Our dataset consists of the following:

- CelebA image dataset
- Blurred CelebA dataset
- Blurred CelebA dataset with differential privacy
- Pixelated CelebA dataset
- Pixelated CelebA dataset with differential privacy

We cropped the images to 128 X 128 centering 128 X 91 and converted to PNG images before we pass as input for training ProGAN model



Dataset example (from left to right):
original, blurred, blurred with DP, pixelized, pixelized with DP

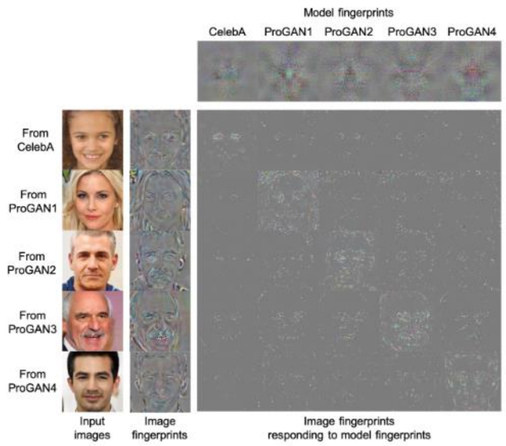
3. Research Challenge

- If image obfuscation techniques can be classified, then the obfuscation can be undone. This renders potentially private data unsafe.
- High Computing power to train huge dataset
- Exposing the GAN fingerprints

4. Algorithms and System Modules

GAN Fingerprints

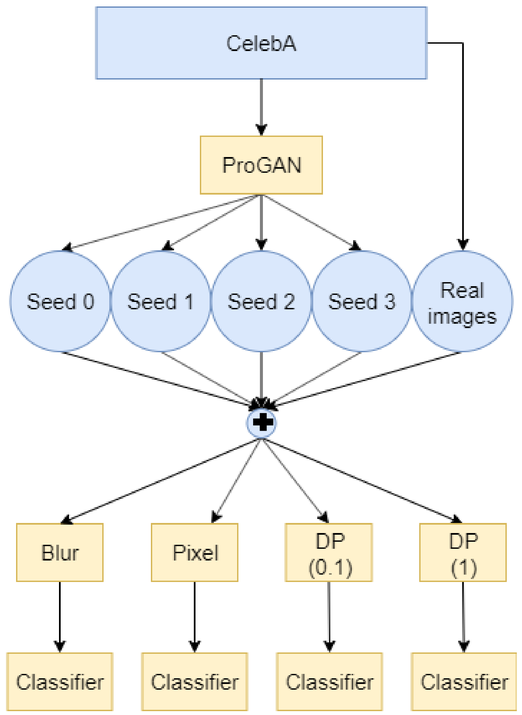
We get to know even a small difference in GAN training (e.g., the difference in initialization) can leave a distinct fingerprint that commonly exists over all its generated images.



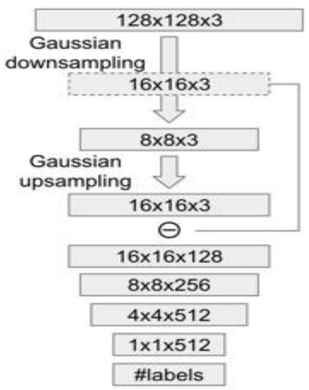
Methodology

- CelebA dataset
  - 200,000 images
- 5 data subsets
  - 10,000 images each
  - 1 subset of real images
  - 4 seeds produced by ProGAN

- Image obfuscation
  - Each set run through each obfuscation method
  - Output fed into classifier
  - The classifier has to predict one of the following labels: ProGAN seed 0-3 and real image



The architecture of the classifier is similar to the architecture of discriminator part of ProGAN.



CLASSIFIER

5. Evaluation

Classification

We evaluated the effectiveness of differential privacy based on the accuracy of the classifier's outputs. The results of our project can be evaluated based on the following concepts:

- If the classifier can more accurately classify blurred/pixelated images than images with differential privacy, then it means that differential privacy is more effective than the current state of the art.
- If the classifier can more accurately classify images with differential privacy than blurred/pixelated images, then the state of the art methods are more effective than differential privacy.

6. Results

Method	Accuracy
Gaussian Blur	80%
Gaussian - Differential Privacy	20%
Pixelization	40%
Pixelization - Differential Privacy	20%

7. References

[1] Yu, Ning, Larry S. Davis, and Mario Fritz. "Attributing fake images to GANs: Learning and analyzing GAN fingerprints." Proceedings of the IEEE International Conference on Computer Vision. 2019.

[2] Fan, Liyue. "Image pixelization with differential privacy." IFIP Annual Conference on Data and Applications Security and Privacy. Springer, Cham, 2018.

8. Acknowledgements

Liyue Fan, University of North Carolina at Charlotte

Stephen Welch, University of North Carolina at Charlotte