Topic: Attribution of Private Images to GANs

Project Stage - 2

Aravind Telidevara
Bhanu Chennam
Sara Ashcraft
Satabdhi Reddy Nalamalpu

Business Partner/Customer Interaction

- Our Computer Vision solution will solve an online privacy problem. This solution will benefit any person who wishes to keep their online images private. The problem is that real images online can be used to create GAN-generated images. This technology can easily be misused. There is already a neural network trained to classify images as real or GAN-generated, but the model's performance on privacy-enhanced images is currently unknown. It is our job to test this.
- What we are trying to solve is to see whether the above mentioned fingerprinting classifiers can adapt to image privacy methods applied to training/testing data. If not, the privacy methods will be shown to provide sufficient protection, or can be utilized by bad actors to pass fake data as authentic.
- We are planning to focus on 4 privacy methods in this project: vanilla pixelization (aka. mosaicing)/blurring functions and their differential privacy versions.
- In short, our objective is:
    - To prevent the classifier from being able to tell which pictures are from which source.
    - To protect images from being unblurred or unpixelated.

- As of now, we did not find any solutions similar to the one that we proposed.

- 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. Specifically, Hidden Markov Models (HMM) can learn to decode redacted documents. Moreover, Convolutional Neural Networks (CNN) are highly adaptable to standard obfuscation, which can re-identify up to 96% of pixelized faces. Therefore, we are trying to improve image obfuscation methods that can provide rigorous privacy guarantees.

- We had met professor Liyue Fan and discussed about the project and its timeline. She has suggested the following things:
    - Read papers about implementing blur, pixelation, etc as a part of protecting the privacy of the GAN generated images.

    - Learn how to pixelate or blur. Check openCV functions for performing these operations on images.
    - Try randomized blur or differential privacy instead.
    - Meet with Dr.Fan at checkpoints to stay on track

Literature Review

Image Pixelization with Differential Privacy by Liyue Fan
(https://link.springer.com/chapter/10.1007/978-3-319-95729-6_10)

- In the paper mentioned above, they have explored how differentially private methods can enhance image obfuscation. The privacy of these differential private methods have been varied and a graph has been plotted between utility and privacy for standard obfuscation techniques and differentially private methods. The result shows that both perform similarly when there is low privacy but there is significant difference when there is high privacy but with low utility.

- As the project we have been working on deals with private images, it is important to know the difference in performances between standard private images and differentially private methods.

- The sensitive content in the images are generally hidden using image obfuscation techniques such as pixelization and gaussian blur. But due to the recent developments in Machine Learning algorithms have made these techniques ineffective for preservation. So we use standard differential privacy notion to image data to enhance standard obfuscation
- There are two differentially private methods described in the paper. They are:
    - Differentially Private Pixelization: The algorithm first performs pixelization on an input image (by computing the average pixel value of each grid cell and applies Laplace perturbation to the pixelated image.
    - Differentially Private Blurring:The initial approach is to first apply Laplace perturbation to each pixel and the apply Gaussian Blur, but it induces noise. To reduce the noise, first apply DP- Pixelization to the input image and the upsample to it the size of original image and then apply Gaussian blur.
- The utility function for obfuscated images by non private methods and differentially private methods is compared . As the degree of privacy is increased for the differentially private methods, the mean square error shows a lower value and a higher value for structural similarity which approaches the utility of non private baselines.

- The re identification attacks through CNN has also been performed on non private as well as the private methods.The non private baselines inflict higher re-identification rates than the differentially private methods which significantly

reduce the attack success rate. The differential blur achieves the lowest re-identification rates which is lower than random guessing which implies that the differential methods are performing well without being traceable.

Practical Image Obfuscation with Provable Privacy by Liyue Fan (https://ieeexplore.ieee.org/abstract/document/8784836)

- In the paper, they proposed a novel image obfuscation solution based on metric privacy, a rigorous privacy notion generalized from differential privacy. The key advantage of our solution is that our privacy model allows for higher utility by providing indistinguishability based on image visual similarity, compared to the current method with standard differential privacy.

- Recently,the principle of differential privacy has been extended and proposed a generalized notion, i.e., metric privacy. Essentially, it defines a distance metric between secrets and guarantees a level of indistinguishability proportional of the distance.Metric privacy guarantees that the output of a mechanism should be roughly the same, i.e., bounded by the distance dX (x, x'), between two inputs x and x'.

- The proposed solution for obfuscated image has 2 major steps: transforming and sampling.A sensitive ROI will first be transformed to a feature vector, and the vector will go through the sampling step to achieve privacy guarantees; the sampled vector will be processed with inverse transform, resulting in the obfuscated ROI image.

Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints by Ning Yu, Larry Davis, and Mario Fritz (https://arxiv.org/abs/1811.08180)

- Recent advances in Generative Adversarial Networks (GANs) have shown increasing success in generating photorealistic images. But they also raise challenges to visual forensics and model attribution. They had presented study of learning GAN fingerprints towards image attribution and using them to classify an image as real or GAN generated. For GAN-generated images, further identifying their sources was also done. The experiments show that
    - GANs carry distinct model fingerprints and leave stable fingerprints in their generated images, which support image attribution
    - Even minor differences in GAN training can result in different fingerprints, which enables fine-grained model authentication
    - Fingerprints persist across different image frequencies and patches and are not biased by GAN artifacts
    - Fingerprint finetuning is effective in immunizing against different types of adversarial image perturbations.
    - Comparisons also show that learned fingerprints consistently outperform several baselines in a variety of setup

Review of Available Open Source Code and Data

- We have found open source codes for image pixelization and image blurring but did not find any open source codes for implementing differential privacy on images.
- We are using CelebFaces Attributes Dataset (CelebA) which is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations.
- Original data and banner image source came from: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
- MMLAB, The Chinese University of Hong Kong is the creator of the data.
- The data is open source but the CelebA dataset is available for non-commercial research purposes only.
- All images of the CelebA dataset are obtained from the Internet which are not property of MMLAB, The Chinese University of Hong Kong and it is not responsible for the content nor the meaning of these images.
- The MMLAB reserves the right to terminate the access to the CelebA dataset at any time.
- Information about the data:
    - 202,599 number of face images of various celebrities
    - 10,177 unique identities, but names of identities are not given
    - 40 binary attribute annotations per image - "1" represents positive while "-1" represents negative
    - 5 landmark locations on the face - left eye, right eye, nose, left mouth, right mouth
    - Bounding box information for each image
    - All the face images, cropped and aligned

Review on Existing Industry Solutions

We have found some research papers addressing the similar problem but with different solutions.

Some of them are:

- GAN based Image Privacy Protection Algorithm: (https://doi.org/10.1117/12.2524274)
    - They had proposed a Generative Adversarial Network (GAN) based image privacy protection algorithm PriGAN, which generates a privacy image for each original image to fool the recognition networks. When generating the privacy image, we consider the technique of adversarial image perturbations (AIP), which could confuse recognition networks with slight perturbations. That is, the privacy image could protect the privacy information by confusing the neural network hosted by the service providers. Meanwhile, the privacy image appears unmodified compared to the original one for human observers, and thus its utility could be preserved.
- Differentially Private Releasing via Deep Generative Model: (https://arxiv.org/pdf/1801.01594.pdf)
    - In this paper, dp-GAN, a general private releasing framework for semantic-rich data was proposed. Instead of sanitizing and then releasing the data, the data curator publishes a deep generative model which is trained using the original data in a differentially private manner; with the generative model, the analyst is able to produce an unlimited amount of synthetic data for arbitrary analysis tasks. In contrast to alternative solutions, dp-GAN highlights a set of key features:
        - it provides theoretical privacy guarantee via enforcing the differential privacy principle
        - it retains desirable utility in the released model, enabling a variety of otherwise impossible analyses
        - it achieves practical training scalability and stability by employing multi-fold optimization strategies
- Privacy-Protective-GAN for Privacy Preserving Face De-Identification: (https://link.springer.com/article/10.1007/s11390-019-1898-8)
    - In this paper, a new framework called Privacy-Protective-GAN (PP-GAN) that adapts GAN (generative adversarial network) with novel verificator and regulator modules specially designed for the face de-identification problem to ensure generating de-identified output with retained structure similarity according to a single input was proposed. The proposed approach in terms of privacy protection, utility preservation, and structure similarity was evaluated.