Deep Obfuscation: Precise Masking of Sensitive Information to Protect Against Machine Learning Adversaries

Yuan Gong and Christian Poellabauer

Department of Computer Science & Engineering, University of Notre Dame

UNIVERSITY OF NOTRE DAME

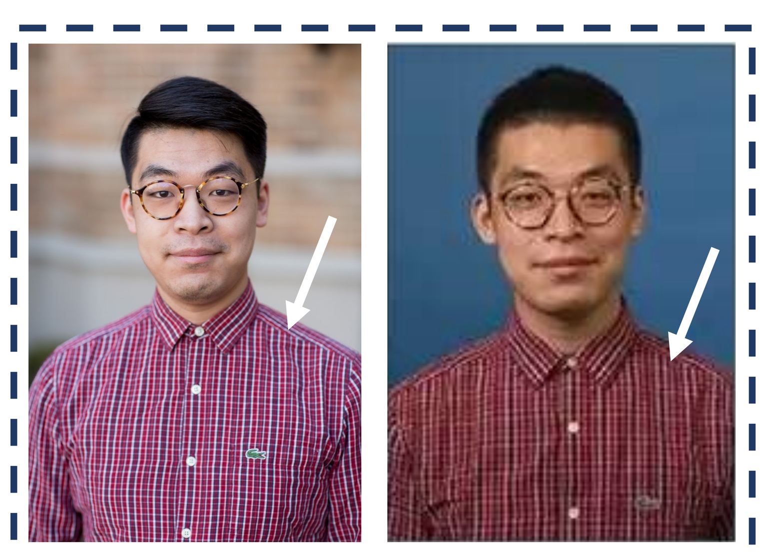
▼ Motivation

Background: It is often desirable to share data (e.g., tables, images, audio, sensor streams, etc.), while removing some sensitive information embedded in such data.

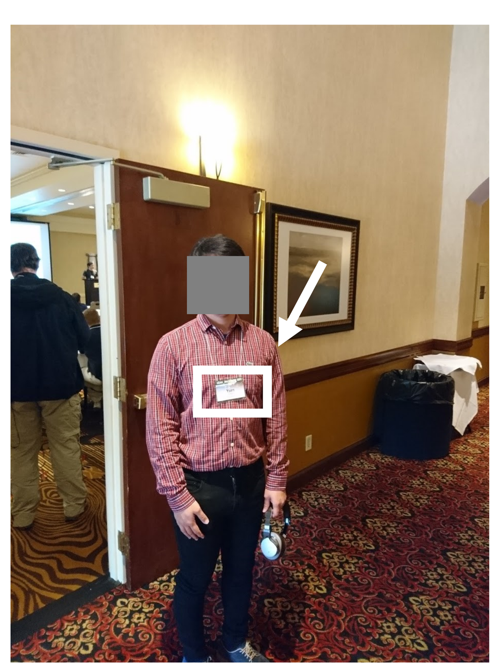
Problem: It is hard to mask data **precisely** and **completely**. A **machine learning adversary** might be able to infer the sensitive information based on **residual cues**.

Solution: We propose ***Deep Obfuscation***, i.e., a deep learning based data masking scheme that aims to **precisely** and **maximally** remove sensitive information from target data, while **minimally** affecting non-sensitive information.

▼ Attack Example

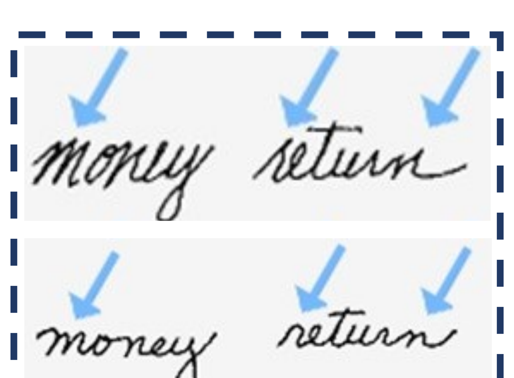


Person A in Training Set



Is this person A?

Blocking face is not sufficient. Clothing, gestures, environment, etc., could be used by an attacker to infer a person's identity.



Handwriting style can reveal the identity, but masking the style also removes the content information.

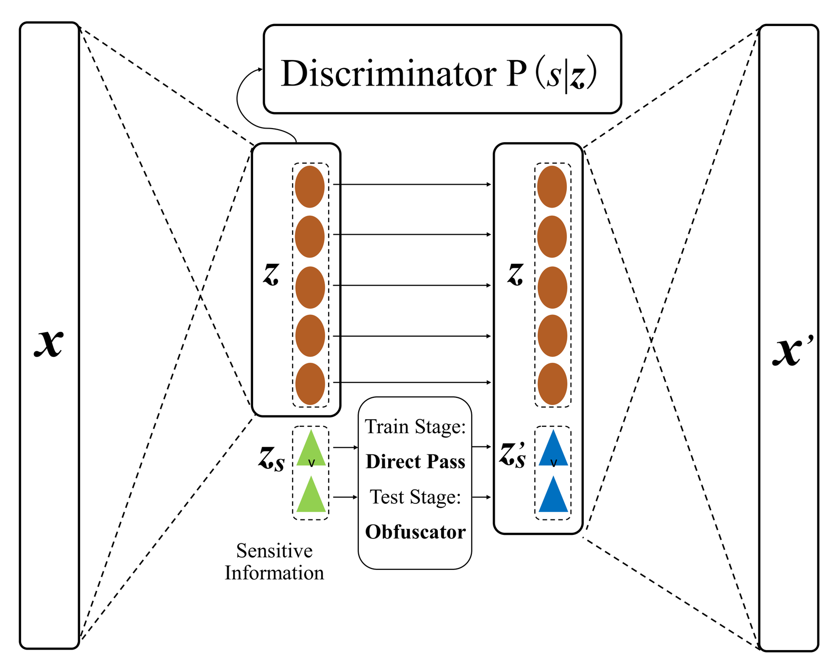
▼ Problem Formalization

Assumption: 1) A data sample $\boldsymbol{x} \sim P_x$ contains some discrete sensitive information s. 2) There exists an underlying relationship $s = F_s(\boldsymbol{x})$.

Attack: An attacker can infer s based on $\boldsymbol{x}$ by learning a machine learning model $P_{model}(s \mid \boldsymbol{x})$ using a dataset $\boldsymbol{D} \sim P_x$ with labels of s .

The Masking Task: Modifying $\boldsymbol{x}$ to $\boldsymbol{x'}$, where $\boldsymbol{x'}$ cannot be used to infer sensitive information, i.e., the inference confidence $P_{model}(s \mid \boldsymbol{x'})$ is small for well-designed machine learning models trained on $\boldsymbol{D} \sim P_{x'}$.

▼ Deep Obfuscation



Data: $\mathcal{D} = \{\{\boldsymbol{x}_1, s_1\}, \{\boldsymbol{x}_2, s_2\}, \ldots, \{\boldsymbol{x}_n, s_n\}\}$
while $\frac{1}{n}\sum_i^n \log P(s_i|\boldsymbol{z}_i) > threshold$ **do**
  for j := 1: # *discriminator_training_epoch* **do**
    optimize discriminator parameters θ_{dis} using Equ. 1;
  end
  for k := 1: # *autoencoder_training_epoch* **do**
    optimize autoencoder parameters θ_{ae} using Equ. 2;
  end
end

- Autoencoder based adversarial training scheme.
- The discriminator serves as the **imaginary adversary**, which tries to estimate sensitive information s from the latent variable $\boldsymbol{z}$, i.e.,
$$\boldsymbol{L}_{\text{discriminator}} = \text{mean}_{\boldsymbol{x} \in \mathbf{D} \sim P_x}(-\log P(s \mid \boldsymbol{z})) \qquad (1)$$
- The goal of the encoder is to avoid such an attack, hence, the discriminator's inference performance is added into the auto-encoder loss, i.e.,
$$\boldsymbol{L}_{\text{autoencoder}} = -\,\boldsymbol{L}_{\text{discriminator}} + \lambda \times \boldsymbol{L}_{\text{reconstruction}} \qquad (2)$$
- The discriminator and the autoencoder are trained alternatively, i.e., during the training of the autoencoder, the parameters of the discriminator are frozen and vice versa.
- This scheme forces the encoder to encode all information other than s in $\boldsymbol{z}$, and the decoder to effectively decode from $\boldsymbol{z}$.

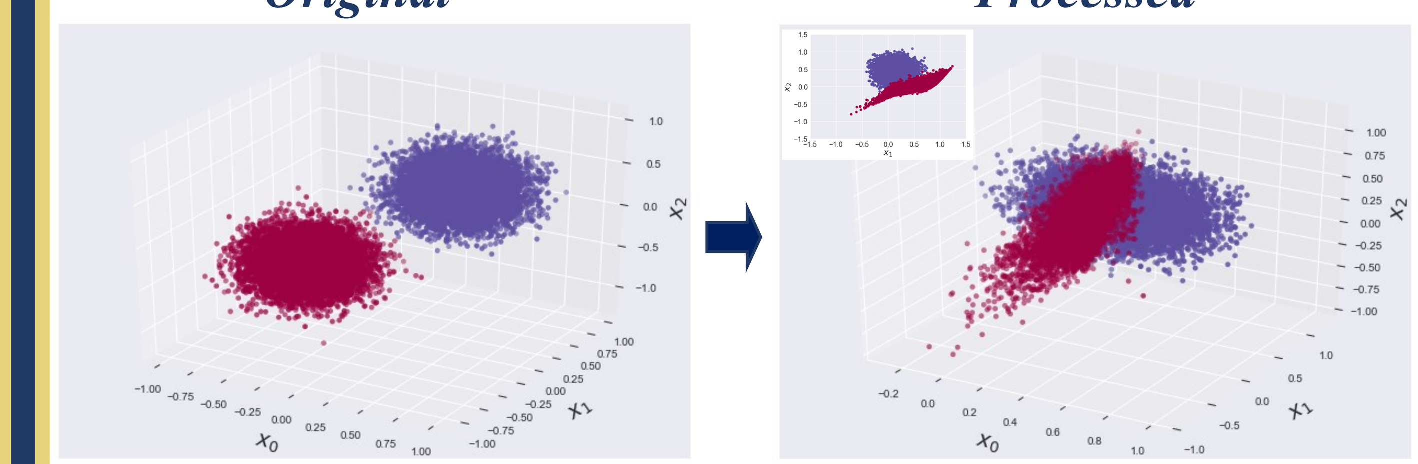
▼ Low Dimensional Example

Task: Remove information of s from $\boldsymbol{x}$

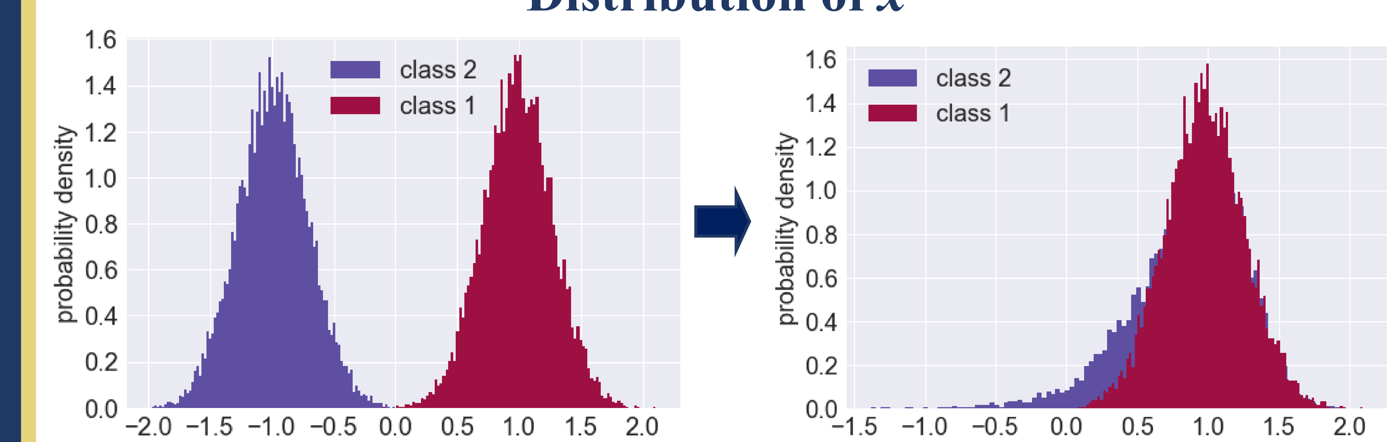
$\boldsymbol{x} \in \mathbf{R}^3$ ~ Mixed Gaussian Distribution

$s = F_s(\boldsymbol{x}) = \mathbf{I}_{w^T x > 0}(\boldsymbol{x})$, $\boldsymbol{w} = [1, 1, 1]$

Original *Processed*

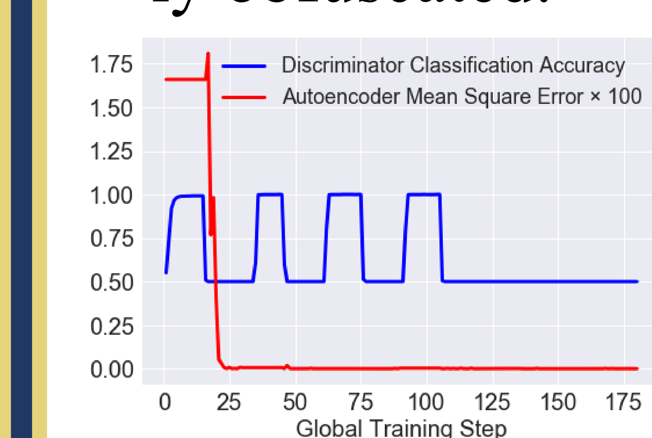


Distribution of $\boldsymbol{x}$



Distribution of $\boldsymbol{w}^T\boldsymbol{x}$

- $\boldsymbol{x'}$ maintains some main characteristics of the original $\boldsymbol{x}$ (e.g., the two Gaussian clusters are still separable).
- Sensitive information is almost completely obfuscated.



- The discriminator and the autoencoder reach a balance.

▼ Future Work

- Design the training procedure and the network architecture for processing high dimensional data (e.g., images) and sequential data (e.g., audio, sensor stream).
- Apply deep obfuscation scheme to a variety type of sensitive information such as identity, health status, and passwords.

Contact: Yuan Gong, ygong1@nd.edu