Improving Robustness to Model Inversion Attacks via Mutual Information Regularization

Tianhao Wang[1], Yuheng Zhang[2], Ruoxi Jia[3]

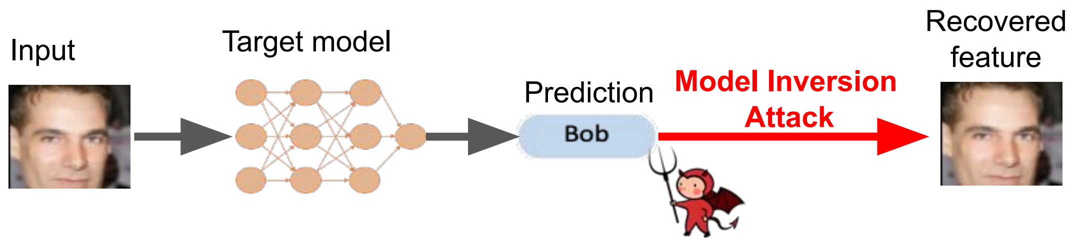
[1]Harvard University [2]University of Illinois Urbana-Champaign [3]Virginia Tech

TL; DR: We formally analyze model inversion attack and propose the **M**utual **I**nformation Regularization based **D**efense (MID).

Motivation

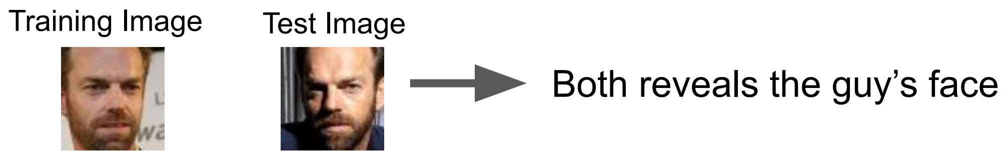
- Existing defense mechanisms against model inversion attack rely on model-specific heuristics or noise injection.
- Existing defense mechanisms significantly hinder model performance.
- We need to design a defense mechanism that is **applicable to a variety of models** and achieves **better utility-privacy tradeoff.**

Model Inversion Attack



Defense Goal

Both the recovery of training images and test images would incur privacy loss to the target identity. We need to design an algorithm to **protect the training data distribution**, instead of just training data set.



MID: Mutual Information Regularization based Defense

Intuition: If the output distribution is independent from input distribution, the attacker cannot learn anything about X's distribution.

Method: Regularize the loss function by the mutual information between model's input and output distribution.

$$\min_{f\in\mathcal{H}} E_{(x,y)\sim p_{X,Y}(x,y)}[\mathcal{L}(y,f(x))] + \lambda \mathcal{I}(X,\hat{Y})$$

Original Loss Function

Regularizer Coefficient

mutual information between input and prediction

$$\mathcal{I}(X,\hat{Y}) = \int_X\int_Y p_{X,Y}(x,y)\log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right)dydx$$

Instantiation of MID

- Linear regression: Taylor-expansion based approximation
- Decision tree: modify ID3
- Deep Neural Networks: information bottleneck technique

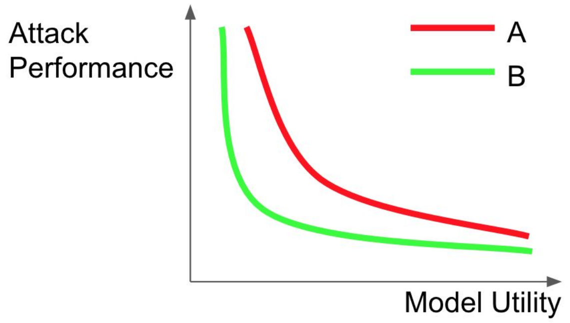
For example, we can regard the neural network as a Markov chain: $Y - X - Z - \hat{Y}$

label　feature　stochastic encoding of intermediate layer　prediction

By Data Processing Inequality, we have $\mathcal{I}(X,\hat{Y}) \le \mathcal{I}(X,Z)$ and we can obtain a new training loss $\min_\theta -\mathcal{I}(Z;Y) + \lambda\mathcal{I}(Z,X)$ which boils down to the **classic information bottleneck**.

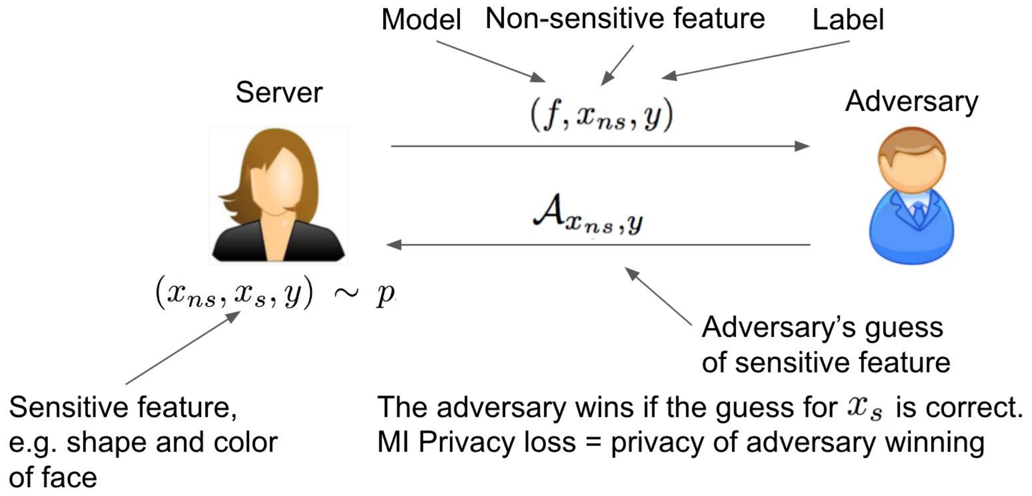
Evaluation Metric

- Defense mechanisms are evaluated in terms of **privacy-utility tradeoff.**
- In the illustration below, the green line is more preferable as it is **more robust against the attack at any fixed model utility.**



Formalizing Model Inversion Attacks

We present a methodology for formalizing model inversion attacks. Unlike previous works that only capture the privacy loss of members in the training set, this is the first attempt of modeling privacy loss of members **in the population.**



Charactering MI Privacy Loss for Differentially Private Models

Main Result: when the learning algorithm is (ε, δ)-differentially private, the MI privacy loss is **tightly** upper bounded by

$$\frac{e^{n\epsilon}-1}{e^{n\epsilon}+1} + \frac{2(e^{n\epsilon}-1)}{(e^{n\epsilon}+1)(e^{\epsilon}-1)}\delta$$

where n is the size of training set. To make bound small, the privacy budget ε needs to be set as **o(1 / #training data)**!

Evaluation on Defending against Various MI Attacks