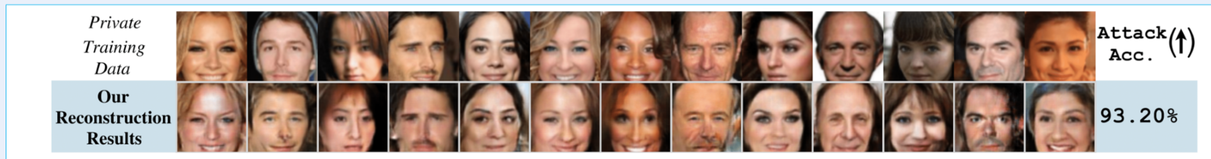
Re-thinking Model Inversion Attacks Against Deep Neural Networks

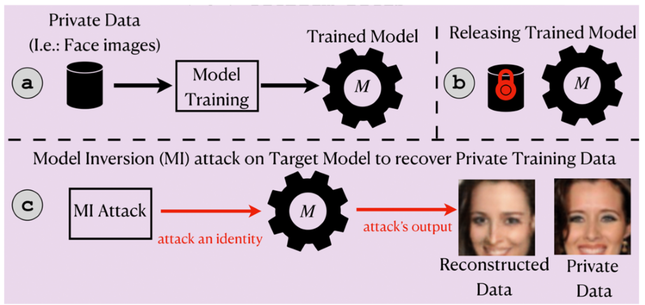
Ngoc-Bao Nguyen(*) Keshigeyan Chandrasegaran(*) Milad Abdollahzadeh Ngai-Man Cheung
thibaongoc_nguyen@mymail.sutd.edu.sg, ngaiman_cheung@sutd.edu.sg
Singapore University of Technology and Design (SUTD)

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA



Model inversion (MI)

Model inversion (MI) attacks aim to infer and reconstruct private training data by abusing access to a model.
MI attacks have raised serious concerns of leaking of sensitive and/or private information (i.e.: face-recognition).



Contributions

- We analyze existing identity loss, argue that it could be sub-optimal for MI, and propose ***an improved identity loss*** that aligns better with the goal of MI.
- We formalize the concept of ***MI overfitting***, analyze its impact on MI and propose a novel solution based on ***model augmentation***. Our idea is inspired by the conventional issue of overfitting in model training and data augmentation as a solution to alleviate the issue.
- We conduct extensive experiments to demonstrate that our solutions can improve SOTA MI algorithms (GMI, KEDMI, VMI) significantly.
- Our solutions achieve for the first time ***over 90%*** attack accuracy under standard CelebA benchmark

Project page & Code & Pretrained models

References

[1] Chen et. al. *Knowledge-enriched distributional model inversion attacks*. In CVPR, 2021.
[2] Zhang et. al. *The secret revealer: Generative model inversion attacks against deep neural networks*. In CVPR, 2020.

* Equal Contribution

SOTA MI Identity loss

The identity loss guides the reconstruction of **x** that is most likely to be recognized by model **M** as identity **y:**

$$L_{id}(x; y = k) = -\log \mathbb{P}_M(y|x) \qquad \text{Eqn. 1}$$

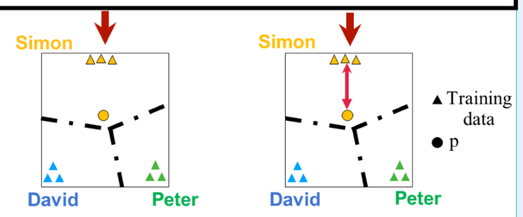
Existing identity loss (Eqn. 1) used in SOTA MI methods **is sub-optimal for model inversion attacks**

Classification (i.e.: recognize between 'Simon', 'Peter' and 'David')

Model Inversion (i.e.: reconstruct data close to private training data of 'Simon')

Eqn. 2 (Existing Work)

$$L_{id}(x; y = k) = -\log \frac{\exp(p^T w_k)}{\exp(p^T w_k) + \sum_{j=1, j\neq k}^{N} \exp(p^T w_j)}$$



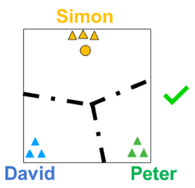
Eqn 2 is suitable as **p** is distant from other classification regions ✓

Eqn 2 is **suboptimal** as **p** is distant from private training data of 'Simon' ✗

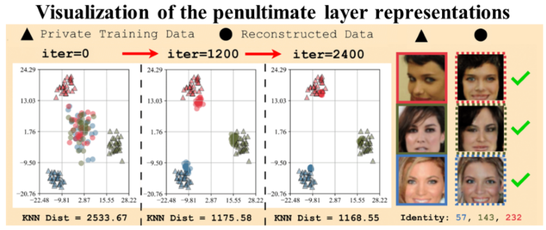
An improved formulation of MI Identity loss

We propose to directly maximize the logit instead of maximizing the log likelihood of class k for MI:

$$L_{id}(x; y = k) = -p^T w_k + \lambda ||p - p_{reg}||_2^2$$



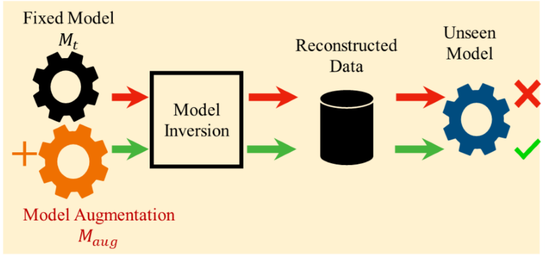
Visualization of the penultimate layer representations



Model Inversion Overfitting

Given the fixed (target) model and our goal of learning reconstructed samples, **we define MI overfitting** as instances which during model inversion, *the reconstructed samples fit too closely to the target model* and adapt to the random variation and noise of the target model parameters, *failing to adequately learn semantics of the identity.*

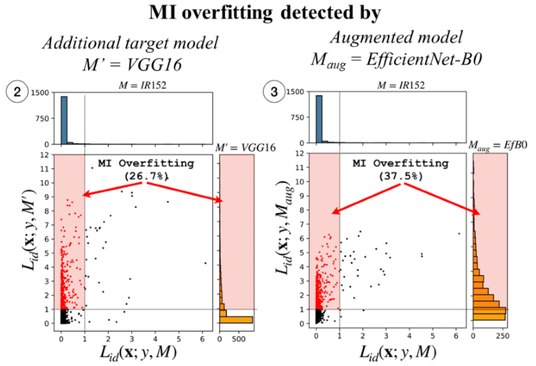
Overcome Model Inversion Overfitting



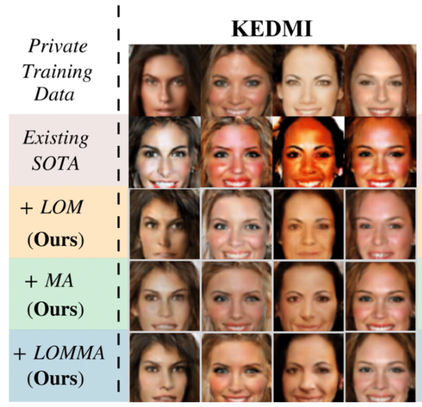
We propose to apply knowledge distillation with target model M_t as the teacher, to train augmented models $M_{aug}^{(i)}$, using the public dataset.

$$L_{id}^{aug}(x; y) = \gamma_t \cdot L_{id}(x; y, M_t) + \gamma_{aug} \cdot \sum_{i=1}^{N_{aug}} L_{id}(x; y, M_{aug}^{(i)})$$

MI overfitting detected by



Main Results



	Attack Acc ↑	KNN Dist ↓
Existing SOTA	80.53%	1247.28
+ LOM **(Ours)**	**92.47%**	**1168.55**
+ MA **(Ours)**	**84.73%**	**1220.23**
+ LOMMA **(Ours)**	**92.93%**	**1138.62**

We report the results for KEDMI and GMI for IR152, face.evoLve and VGG16 target model. Following exact experiment setups in [1], here Dpriv = CelebA, Dpub = CelebA, evaluation model = face.evoLve.

Method	Attack Acc ↑	Imp. ↑	KNN Dist ↓
CelebA/CelebA/IR152			
KEDMI	80.53 ± 3.86	-	1247.28
+ LOM (Ours)	92.47 ± 1.41	11.94	1168.55
+ MA (Ours)	84.73 ± 3.76	4.20	1220.23
+ LOMMA (Ours)	**92.93 ± 1.15**	**12.40**	**1138.62**
GMI	30.60 ± 6.54	-	1609.29
+ LOM (Ours)	78.53 ± 3.41	47.93	1289.62
+ MA (Ours)	61.20 ± 4.34	30.60	1389.99
+ LOMMA (Ours)	**82.40± 4.37**	**51.80**	**1254.32**
CelebA/CelebA/face.evoLve			
KEDMI	81.40 ± 3.25	-	1248.32
+ LOM (Ours)	92.53 ± 1.51	11.13	1183.76
+ MA (Ours)	85.07 ± 2.71	3.67	1222.02
+ LOMMA (Ours)	**93.20 ± 0.85**	**11.80**	**1154.32**
GMI	27.07 ± 6.72	-	1635.87
+ LOM (Ours)	61.67 ± 4.92	34.60	1405.35
+ MA (Ours)	74.13 ± 4.32	47.06	1352.25
+ LOMMA (Ours)	**82.33 ± 3.51**	**55.26**	**1257.50**
CelebA/CelebA/VGG16			
KEDMI	74.00 ± 3.10	-	1289.88
+ LOM (Ours)	89.07 ± 1.46	15.07	1218.46
+ MA (Ours)	82.00 ± 3.85	8.00	1248.33
+ LOMMA (Ours)	**90.27 ± 1.36**	**16.27**	**1147.41**
GMI	19.07 ± 4.47	-	1715.60
+ LOM (Ours)	69.67 ± 4.80	50.60	1363.81
+ MA (Ours)	51.73 ± 6.03	32.66	1467.68
+ LOMMA (Ours)	**77.60 ± 4.64**	**58.53**	**1296.26**