Label-Only Model Inversion Attacks via Knowledge Transfer

Ngoc-Bao Nguyen(*) Keshigeyan Chandrasegaran(*) Milad Abdollahzadeh Ngai-Man Cheung

thibaongoc_nguyen@mymail.sutd.edu.sg, ngaiman_cheung@sutd.edu.sg

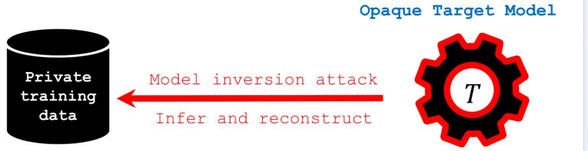
(*) Equal contribution

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

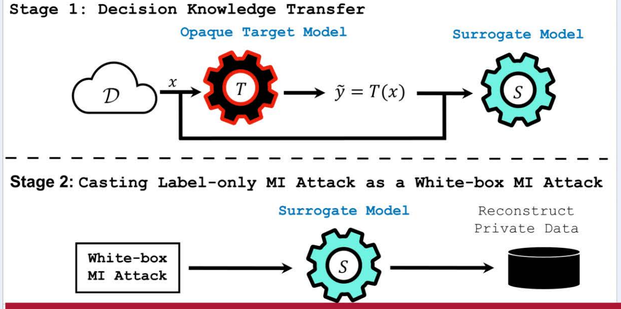
NEURAL INFORMATION PROCESSING SYSTEMS

Label-only Model Inversion (MI)

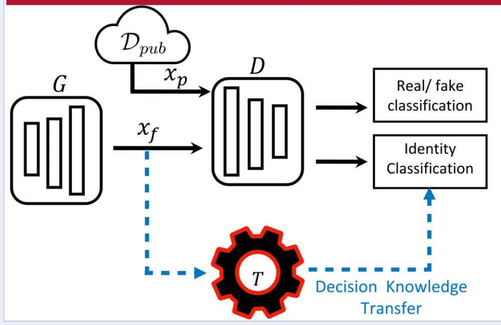
Label-only Model inversion (MI) attacks aim to infer and reconstruct private training data by abusing access to a model's predicted label (hard label) without confidence scores nor any other model information.



Label-only MI via Knowledge Transfer (LOKT)



Decision Knowledge Transfer using our Target model-assisted ACGAN



We propose a new T-ACGAN to leverage generative modeling and the target model for effective knowledge transfer.

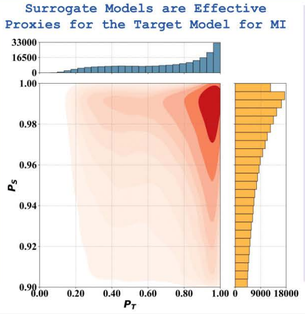
$$\mathcal{L}_{D,C} = -E[\log P(s = Fake|x_f)] - E[\log P(s = Real|x_p)] - E[\log P(c = \tilde{y}|x_f)]$$

With T-ACGAN capturing the data manifold of public samples, synthetic data is diverse and abundant. We hypothesize that such rich synthetic data could lead to improved decision knowledge transfer.

Analysis for justification of surrogate models

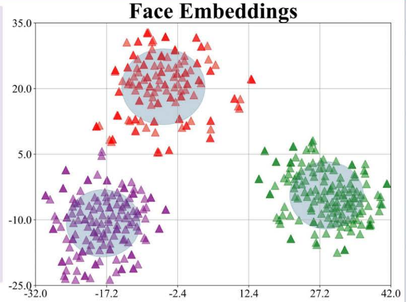
Property P1:
For high-likelihood samples under S, it is likely that they also have high likelihood under T.

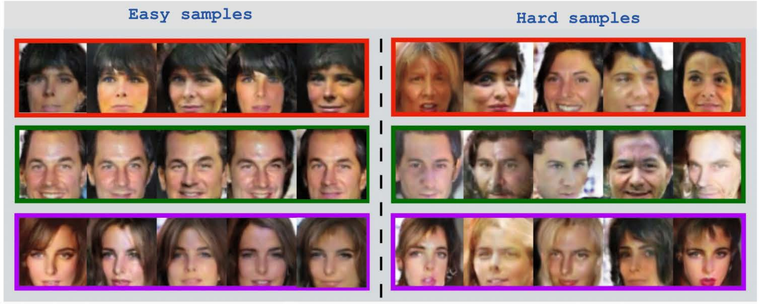
P1 is essential for performing MI attacks on S because during the inversion, we are finding samples that have high probability under T.

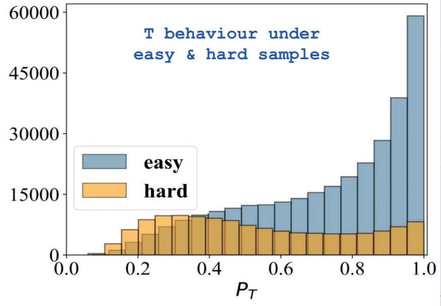


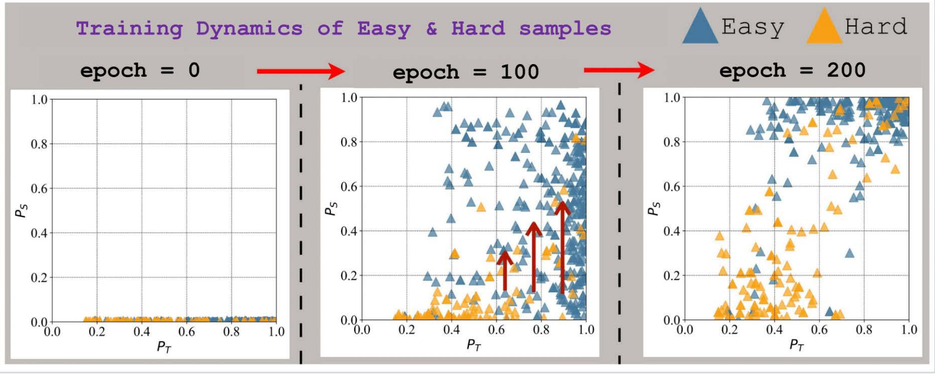
The general learning dynamics of DNNs:

- "Easy samples" are ones that fit better some patterns in the data (and correspondingly "hard samples").
- In DNNs learning, the models learn simple and general patterns of the data first in the training stage to fit the easy samples.









Main Results

Setup	Attack		Attack acc. ↑	KNN dt. ↓
T = FaceNet64	BREPMI		73.93 ± 4.98	1284.41
$\mathcal{D}_{priv}$ = CelebA	LOKT	$C \circ D$	81.00 ± 4.79	1298.63
$\mathcal{D}_{pub}$ = CelebA		S	92.80 ± 2.59	1207.25
		S_{en}	**93.93 ± 2.78**	**1181.72**
T = IR152	BREPMI		71.47 ± 5.32	1277.23
$\mathcal{D}_{priv}$ = CelebA	LOKT	$C \circ D$	72.07 ± 4.03	1358.94
$\mathcal{D}_{pub}$ = CelebA		S	89.80 ± 2.33	1220.00
		S_{en}	**92.13 ± 2.06**	**1206.78**
T = VGG16	BREPMI		57.40 ± 4.92	1376.94
$\mathcal{D}_{priv}$ = CelebA	LOKT	$C \circ D$	71.33 ± 4.39	1364.47
$\mathcal{D}_{pub}$ = CelebA		S	85.60 ± 3.03	1252.09
		S_{en}	**87.27 ± 1.97**	**1246.71**
T = FaceNet64	BREPMI		43.00 ± 5.14	1470.55
$\mathcal{D}_{priv}$ = CelebA	LOKT	$C \circ D$	43.27 ± 3.53	1516.18
$\mathcal{D}_{pub}$ = FFHQ		S	59.13 ± 2.77	1437.86
		S_{en}	**62.07 ± 3.89**	**1428.04**
T = BiDO	BREPMI		37.40 ± 3.66	1500.45
$\mathcal{D}_{priv}$ = CelebA	LOKT	$C \circ D$	45.73 ± 5.94	1493.48
$\mathcal{D}_{pub}$ = CelebA		S	58.53 ± 4.87	1427.22
		S_{en}	**60.73 ± 3.07**	**1395.93**

Contributions

- We propose a new approach for **Label-Only MI attack using Knowledge Transfer (LOKT)** by transferring decision knowledge from the target model to surrogate models and performing white-box attacks on the surrogate models. Our proposed approach is the first to address label-only MI via white-box attacks.
- We propose a new **Target model-assisted ACGAN (T-ACGAN)** to leverage generative modeling and the target model for effective knowledge transfer.
- We perform **analysis** to support that our surrogate models are effective proxies for the target model for MI.