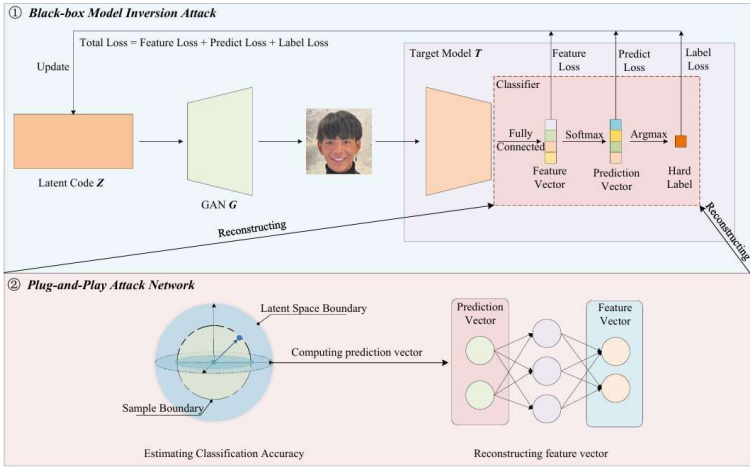
From Black-box to Label-only: a Plug-and-Play Attack Network for Model Inversion

Huan Bao, Kaimin Wei†, Yao Chen, Hanting Hou, Jinpeng Chen,Yongdong Wu†.
College of Cyber Security, Jinan University, China
School of Computer Science, Beijing University of Posts and Telecommunications, China

BMVC 2024

Introduction

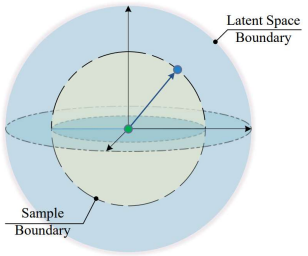
Model inversion (MI) attacks can reconstruct the private training data of deep neural networks. Nevertheless, existing black-box MI attacks significantly depend on the soft labels obtained from the classifier of the target model, which restricts their practicality.



In this paper, we present a generic Plug & Play Attack Network (PnPAN) for MI, which is the first general framework to transform black-box MI attacks into hard-label-only ones. The fundamental idea of this framework is to assess the existing latent code exclusively using hard labels and employ a pre-trained reverse network. This might make it possible to reconstruct the classifier of the target model with just hard labels. Extensive experimental results demonstrate our approach's performance superiority compared with the state-of-the-art label-only attack and its broad applicability.

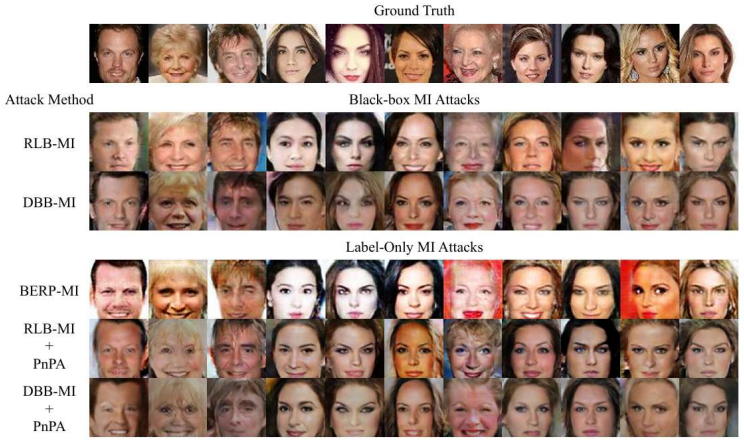
Plug & Play Attack Network

For a specific category, it can be mapped to a point in the latent space of the GAN. If this point sufficiently represents the category, it should be located near the center of that category. Hence, sampling around this point should yield samples belonging to the same category. Extending this idea further, if the point is near the boundary of the category, sampling around it and calculating the proportion of samples that match the category can, to some extent, represent the accuracy of this point's membership in that category.



By training a classification model on public datasets as an additional feature extractor, a reverse network model is pre-trained to reconstruct predicted accuracy into feature vectors. This approach more effectively leverages the classification results of the target model to enhance the reconstruction performance.

Visualization of Reconstructed Results

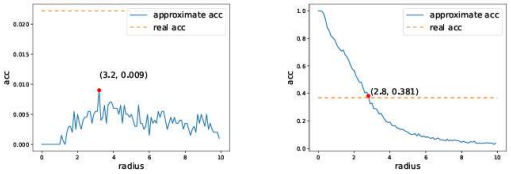


Experiment

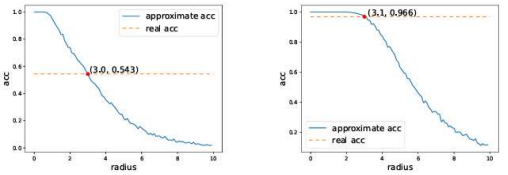
Results in Various Structures

Model	Type	Method	ACC↑	PSNR↑	KNN Dist↓
VGG16	Black-box	RLB-MI	0.642	15.80	1262.34
		DBB-MI	0.858	20.66	1180.63
	Label-only	BERP-MI	0.562	13.36	1872.48
		RLB-MI+PnPA	0.490	15.31	1421.46
		DBB-MI+PnPA	**0.830**	**18.74**	**1253.15**
FaceNet64	Black-box	RLB-MI	0.804	16.27	1354.86
		DBB-MI	0.916	18.06	1091.15
	Label-only	BERP-MI	0.734	13.69	1685.29
		RLB-MI+PnPA	0.650	15.42	1398.40
		DBB-MI+PnPA	**0.900**	**16.63**	**1106.87**
ResNet-152	Black-box	RLB-MI	0.812	15.23	1308.69
		DBB-MI	0.898	17.37	1063.38
	Label-only	BERP-MI	0.754	13.17	1745.73
		RLB-MI+PnPA	0.660	14.91	1384.12
		DBB-MI+PnPA	**0.870**	**16.52**	**1165.72**

The impactof sampling radius



(a) real acc = 0.020 (b) real acc = 0.381



(c) real acc = 0.543 (d) real acc = 0.966