Reinforcement Learning-Based Black-Box Model Inversion Attacks

Gyojin Han Jaehyun Choi Haeil Lee Junmo Kim
School of Electrical Engineering, KAIST
{hangj0820, chlwogus, haeil.lee, junmo.kim}@kaist.ac.kr

A. Network architectures of Soft Actor-Critic

In this section, we describe the architectures of the actor networks and critic networks of Soft Actor-Critic (SAC) [4] agents used in RLB-MI. Table 1 shows the hyperparameters for both networks.

Parameter	Value
Number of hidden layers	2
Number of hidden units per layer	256
Activation function	ReLU

Table 1. Network architectures of both actor networks and critic networks.

B. Experiments on digit classification task

To evaluate our method on a task other than face recognition, we experiment with the baseline model inversion attacks [1,5,8,9] and RLB-MI on the digit classification task. We use a network with 3 convolutional layers and 2 pooling layers as a target model and a network with 5 convolutional layers and 2 pooling layers as an evaluation model. We train the target model with a private dataset, MNIST handwritten digit data [6], and use EMNIST-Letters [3] as a public dataset. We reconstruct 10 images of each digit from 0 to 9 using 10 random seeds. Therefore, all attacks are evaluated with a total of 100 generated images. Our proposed method, RLB-MI, outperforms other baseline attacks as shown in Table 2. In addition, Figure 1 shows that RLB-MI reconstructs the important features of each digit.

Type	Method	Attack Acc	KNN Dist	Feat Dist
White-box	GMI	0.840	10.4	28.7
	KED-MI	0.980	22.0	59.1
Black-box	LB-MI	0.400	30.1	78.9
	RLB-MI (Ours)	**1.000**	**6.3**	**23.1**
Label-only	BREP-MI	0.960	12.1	39.9

Table 2. Attack performance of the model inversion attacks on the MNIST digit classifier.

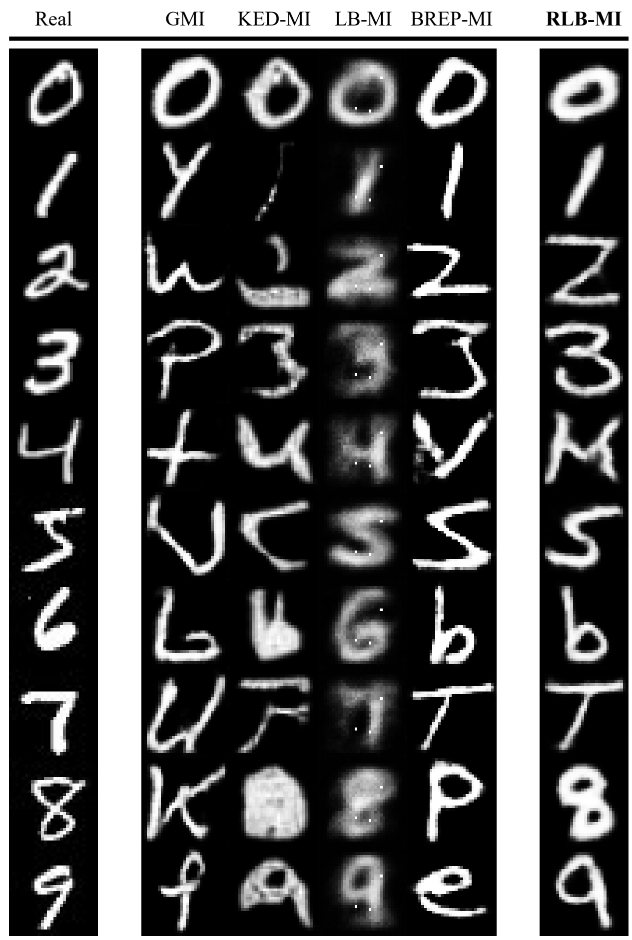

Figure 1. Real samples and attack images from MNIST. The left-most images are real samples and the images in the same row are for the same digit.

C. Ablation Study on Reward Term

To better understand the effect of the proposed reward term r_3 on the performance of our attack, we conduct an ablation study on r_3. The performance of our attack is evaluated with and without the r_3 term in the reward function.

We use the target model of Face.evoLVe [2] architecture trained with CelebA [7] for this experiment. As shown in Table 3, it can be observed through all the metrics that the attack performance is significantly improved when r_3 is included in the reward function.

Method	Attack Acc	KNN Dist	Feat Dist
RLB-MI (w/o r_3)	0.576	1361.5	1273.7
RLB-MI (with r_3)	0.793	1225.6	1112.1

Table 3. Ablation study results on the reward term r_3.

References

[1] Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched distributional model inversion attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16178–16187, October 2021. 1

[2] Yu Cheng, Jian Zhao, Zhecan Wang, Yan Xu, Karlekar Jayashree, Shengmei Shen, and Jiashi Feng. Know you at one glance: A compact vector representation for low-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017. 2

[3] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926, 2017. 1

[4] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018. 1

[5] Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15045–15053, June 2022. 1

[6] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The mnist database of handwritten digits. 1998. 1

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

[8] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pages 225–240, New York, NY, USA, 2019. ACM. 1

[9] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1