SPARSE BLACK-BOX INVERSION ATTACK WITH LIMITED INFORMATION

Yixiao Xu, Xiaolei Liu , Teng Hu, Bangzhou Xin, Run Yang*

Institute of Computer Application
China Academy of Engineering Physics, China

ABSTRACT

Existing black-box model inversion attacks mainly focus on training and attacking surrogate models. However, due to the deployment process of face recognition models, training surrogate models becomes extremely difficult in practice. At the same time, query-based black-box inversion attacks still suffer from low image quality and high computational costs. To bridge these gaps, in this paper, we propose BMI-S, a sparse black-box inversion attack against face recognition models. BMI-S first introduces evolution strategies to perform efficient black-box gradient estimation and achieve query-based attacks. Meanwhile, BMI-S performs sparse attacks on the key styles that contribute most to the face recognition process. By only optimizing key style control vectors, BMI-S further narrows the dimensions of the search space and accelerates the inversion attacks.

***Index Terms*—** model inversion, deep neural network, data privacy

1. INTRODUCTION

Recent studies find that deep learning models are vulnerable to model inversion attacks [1, 2, 3, 4]. Guided by the output of the victim model, malicious users can reconstruct privacy-sensitive characteristics of the training set, which leads to a great threat of privacy leakage. Fredrikson et al. [1, 2] first achieve white-box model inversion attacks on linear regression models and shallow neural networks, using a gradient-based approach. However, this method fails to provide a meaningful result when dealing with more complex deep neural networks. Therefore, some later researches introduce the Generative Adversarial Network (GAN) into the model inversion process and improve the quality of reversed examples [3, 5, 6].

Limitations of previous methods. Despite the success of white-box inversion attacks, existing efforts under black-box constraints are still faced with the following challenges:

(C1) Most black-box inversion attacks are based on training and attacking a surrogate model [7, 8, 9], which requires additional training costs. Moreover, when focusing on face recognition tasks, the deployment process of face recognition models may cause it difficult to train surrogate models (which we will further illustrate in Section 3).

(C2) The query-based method [4] suffer from high computational costs. Thus the required query numbers and the quality of reversed images are unacceptable.

To address the above-mentioned challenges, in this paper, we propose BMI-S, a sparse black-box model inversion attack against face recognition models. Taking advantage of evolution strategies, we replace the gradient calculation process with black-box gradient estimation in GAN-based white-box attacks [3, 5, 6], and expand the GAN-based method to black-box constraints. Furthermore, we explore the semantic relationship of the GAN's input and the features of reversed images, then sparsely optimize the key styles that are critical to the recognition process. Our main contributions can be summarized as follows:

- We propose an efficient black-box model inversion attack method: BMI-S. The proposed method is query-based, thus it is applicable for face recognition conditions where training surrogate models is difficult (addressing C1).
- We perform sparse attacks based on styles, which narrows the search space. The reversed images of our method have a high quality and the attack process is significantly accelerated. (addressing C2).

2. RELATED WORK

White-box model inversion attacks. Model inversion attacks (MIAs) aim to reconstruct the privacy features of the training data. Fredrikson et al. [1] first implemented model inversion attacks on linear regression models using a gradient-based approach [2]. However, these initial attacks fail to produce meaningful results when processing deep neural networks (DNNs) with more complex architectures. To perform model inversion attacks on DNNs, Zhang et al. [3] introduce the Generative Adversarial Network (GAN) to reduce the search space, the GAN-based model inversion attacks (GMI) provide results with meaningful semantic information. Recent researches mainly follow the GMI [3] and

*Correspondence should be addressed to Xiaolei Liu, luxaole@gmail.com. This research was supported by the National Natural Science Foundation of China (Grant No. 62102379 and No. U20B2046).

attempt to improve it from different perspectives. Struppek et al. [5] apply random augmentations to the intermediate results to enhance the robustness of the attacks. Khosravy et al. [6] conclude that the sparse distribution of features in the search space limits the attack efficiency, thus they use VAE-GAN [10] to overcome the tackle.

Black-box model inversion attacks. Black-box model inversion attacks can be mainly divided into transfer-based attacks and query-based attacks. Existing methods mainly explore the transfer-based attack [7, 8, 9]. In transfer-based attacks, adversaries first train a surrogate model with a similar performance to the target model and then perform white-box attacks on the surrogate model to get inversion results. Mehnaz et al. [4] propose the first query-based black-box model inversion attack. However, the time complexity of their method is linearly and positively related to the dimensions of the data, making it less practicable when dealing with high-dimensional data.

3. METHODOLOGY

3.1. Problem Formulation

Deep face recognition (training). Denote the target face recognition model as F, a benign face example as $\boldsymbol{x}$. The training process of F is to build up a mapping $F : \boldsymbol{x} \rightarrow \boldsymbol{y}$, where $\boldsymbol{y} = \{y_0, y_1, ..., y_n | y \in [0, 1]\}$ represents the prediction result of N categories.

Deep face recognition (deployment). In practice, face recognition models are usually not used to recognize faces in the training set. The deployment of a trained face recognition model can be divided into two steps: (1) Face registration. The model owner uses the trained face recognition model to extract and store the features of several registered faces (e.g. faces of a company's employers). (2) Face inference. When fed with an input image, the model first extracts its features and compares the features with those of registered faces, then makes predictions based on the similarity of features.

Model inversion attack. For MIA, the adversary's goal is to reconstruct a synthetic image $\hat{x}_k$ which contains privacy features of the target class k [5]. A straightforward way is to define the process as the following optimization problem:

$$\hat{x}_k = \underset{\hat{x}_k}{argmin}\ L(F, \hat{x}_k, k) \tag{1}$$

where L denotes the classification loss (e.g. cross-entropy loss). However, since semantic-meaningful images are sparsely distributed in the feature space [6], Equation 1 often falls into local minima and fails to produce meaningful results.

Recent MIA methods [3, 5, 6] introduce the GAN structure to narrow the search space and generate semantic-meaningful images. Denote the GAN as G and the input vector of the GAN as $\boldsymbol{z}$, then the optimization problem in Equation 1 can be reformulated as:

$$\boldsymbol{z} = \underset{\boldsymbol{z}}{argmin}\ L(F, G(\boldsymbol{z}), k) \tag{2}$$

where the search space of $\boldsymbol{z}$ is much smaller than $\hat{x}_k$ in Equation 1.

For black-box conditions, recent researches provide two approaches to solve the optimization problem defined in Equation 2: transfer-based attack [7, 8, 9] and query-based attack [4]. However, it is difficult to train a surrogate model for face recognition tasks in practice.

Why it is difficult to train a surrogate model for face recognition tasks in practice? Training a surrogate model is to mimic the behaviors of the teacher model towards different inputs. However, according to the deployment of face recognition models, the outputs of deployed models are narrowed to a much smaller space, which makes it difficult for the surrogate model to learn from the teacher model. Therefore, it is more practicable to perform query-based attacks on face recognition models.

3.2. Black-box Model Gradient Estimation

Some previous researches implement gradient-estimation-based adversarial attacks under black-box conditions [11, 12, 13]. The proposed methods show that the Natural Evolution Strategy (NES) [14] is a cost-acceptable way to estimate gradients. We also use NES as the black-box gradient estimator in BMI-S. Combining with Equation 2, a one-step NES gradient estimation can be expressed as follow:

$$\boldsymbol{\hat{g}_z} = \frac{1}{m} \sum_{i=1}^{m} L(F, G(\boldsymbol{z} + \mu \boldsymbol{\delta_i}), k) \cdot \boldsymbol{\delta_i} \tag{3}$$

where $\boldsymbol{\hat{g}_z}$ denotes the estimated gradient and μ denotes a small smooth parameter (e.g. 0.001). The NES estimator first randomly initialize m vectors $\{\delta_0, \delta_1, ..., \delta_m\}$, representing potential gradient directions. Then the estimator use the loss function L to evaluate each random direction. Finally, the estimated gradient is calculated as a weighted average of m random directions.

After the one-step gradient estimation, BMI-S update the input vector z as:

$$\boldsymbol{z_{i+1}} = \boldsymbol{z_i} + lr \cdot \boldsymbol{\hat{g}_z} \tag{4}$$

where i denotes the number of current iterations and lr denotes the learning rate of the optimization process.

For adversarial attacks, an imperfect gradient estimator is sufficient to perform a successful attack [12]. However, the original NES-based gradient estimator often gets stuck in local minima when performing model inversion attacks because the distribution of the target $\boldsymbol{z}$ in the search space is narrow. Therefore, we further improve the gradient estimation process by performing sparse attacks.

3.3. Style-based Sparse Attack

Following the Plug&Play Attack [5], we choose StyleGAN2 [15] as the image generator. The image generation process of StyleGAN2 is a two-step process as follow:

$$\boldsymbol{w} = \{\boldsymbol{w_1}, \boldsymbol{w_2}, ..., \boldsymbol{w_q}\} = M(\boldsymbol{z}) \tag{5}$$

the model first use a mapping function M to map the input vector $\boldsymbol{z}$ into q style vectors: $\boldsymbol{w} = \{\boldsymbol{w_1}, \boldsymbol{w_2}, ..., \boldsymbol{w_q}\}$. These vectors are transport to different layers of the generator and control different characteristics of generated images.

$$\boldsymbol{x} = S(\boldsymbol{w}, \boldsymbol{n}), \tag{6}$$

Then the model use the generate function S to construct the final image $\boldsymbol{x}$, where $\boldsymbol{n}$ denotes the random noise witch controls the details of images (e.g., skin texture). Therefore, the optimization problem in Equation 2 changes to:

$$\boldsymbol{w} = \underset{\boldsymbol{w}}{argmin}\ L(F, S(\boldsymbol{w}, \boldsymbol{n}), k) \tag{7}$$

According to the evaluation of StyleGAN2 [15], the author find that style vectors $\boldsymbol{w}$ transported to different layers control different characteristics. Thus we consider to choose style vectors that have major contribution to the evaluation function F. By only updating these key style vectors and leaving others unchanged, we can further narrow the search space and accelarate the optimization problem.

Intuitively, some identity features (e.g., the eye and nose) of a face image are more important than some others (e.g. the hairstyle). By searching for the key style vectors that control the more important identity features, BMI-S sufficiently narrows the dimensions of search space.

4. EXPERIMENTS

4.1. Experimental Settings

Dataset. We mainly use two face recognition datasets to evaluate our method, CelebA [16] and MUCT [17].
Model. For face recognition models, we use the open-source pre-trained IR-50 [18] and IR-101 [18] provided by TFace[1]. As for GAN-based face generation, we use StyleGAN2 model pre-trained on the FFHQ dataset provided by [15].
Baseline Method. We fine-tune the query-based inversion attack method in [4] to face recognition tasks as a baseline method. Since it has a similar attack process like FD attack [11] which is originally used to perform adversarial attacks, we use FD to represent the fine-tuned method.
Metrics. We evaluate our method using three metrics. (1) Success Rate (SR): the percentage of results that satisfy the success condition. (2) Fréchet Inception Distance inception distance (FID) [19]: the feature distance score used to evaluate images generated by GANs. (3) Confidence Score (CS): the similarity score of origin and reversed images produced by the online face comparison model[2].
Attack Settings. For StyleGAN2, we set the number of style control vectors as 4, and the input dimensions as 512. For each attack, the maximum query times is $30,000$. The success condition is that the reversed feature vector have a cosine similarity score $s > 0.75$ with the registered feature vector. The population size of NES is 20. We implemented BMI and BMI-S on the Pytorch [20] platform.

4.2. Transfer-based Attacks

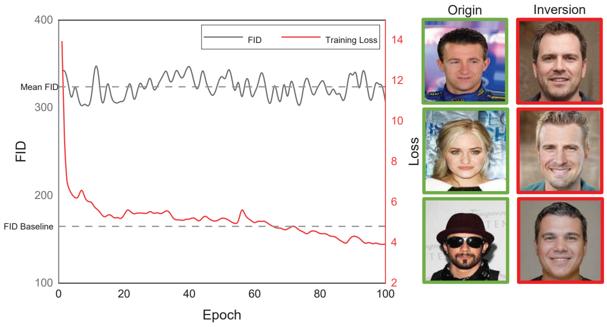

Fig. 1. Evaluation results of transfer-based attacks. For each epoch, we perform GAN-based white-box model inversion attacks on the surrogate model and calculate the FID score. The FID baseline is calculated between real images. We give three pairs of original and reversed images on the right side by attacking the final fixed surrogate model.

We fist perform some experiments to illustrate it is difficult to train a surrogate model for a deployed face recognition model. We use a pre-trained IR-101 model as the teacher model and use another untrained IR-50 model as the student model. Then we randomly choose 10 images from CelebA as the registered faces. For each input image, the teacher model will output the similarity scores of the input image with the registered ones. Thus the goal of the student model is to output similar scores like the teacher model when given a test image. We train the student model for 100 epochs and record the mean loss of each epoch. For each epoch, we perform a white-box model inversion attack using the Plug&Play attack [5] to get several reversed images and calculate the FID score using original and reversed images.

According to Fig 1, after 100 epochs of training, the training loss become stable. However, the reversed images are still far different from the original ones from both FID and visual perspectives, which proves that the student model fail to learn a similar feature extraction ability from the teacher model. Therefore, performing transfer-based inversion attacks on a deployed face recognition model will be difficult.

[1] https://github.com/Tencent/TFace/tree/master/recognition

[2] https://cloud.tencent.com/product/facerecognition

Table 1. Performance comparison of three query-based black-box model inversion attacks against different dataset and target models. The evaluation metrics include Success Rate (SR), Frechet inception distance (FID), and Confidence Score (CS).

Dataset	Model	Method	SR↑	FID↓	CS↑
CelebA	IR-50	FD	0.09	192.50	0.0364
		BMI	0.79	150.83	0.7211
		BMI-S	**0.86**	**105.08**	**0.7518**
	IR-101	FD	0.07	170.94	0.0485
		BMI	0.81	138.76	0.7449
		BMI-S	**0.88**	**98.13**	**0.7913**
MUCT	IR-50	FD	0.05	237.35	0.0264
		BMI	0.74	187.78	0.6236
		BMI-S	**0.82**	**129.11**	**0.7148**
	IR-101	FD	0.03	220.50	0.0297
		BMI	0.76	174.68	0.6864
		BMI-S	**0.84**	**130.06**	**0.7115**

4.3. Performance Comparison

We then list the overall performance of different query-based methods in Tab 1. According to Tab 1, we have following observations: (1) FD fails to produce meaningful results with acceptable queries because it requires 1024 queries for each gradient estimation step, while the number is 20 for BMI and BMI-S. (2) Compared with BMI, BMI-S achieves a higher success rate and lower FID score due to the smaller search space narrowed by sparse attack. (3) BMI and BMI-S have better performances on IR-101 than IR-50, which indicates that the feature extraction ability of the target model will also affect the quality of reversed images.

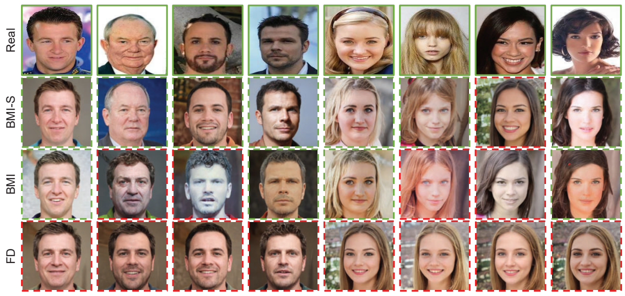

Fig. 2. Results of BMI-S, BMI, and FD attacks against IR-50 models. Registered images are randomly selected from CelebA dataset. Reversed images with green borders indicate that they can pass the online face comparison service, while those with red borders fail to.

Fig 2 gives a visual example of the results generated by different query-based attacks. It can be observed that BMI-S achieves the highest similarity compared with the other two methods, and most results of BMI-S can pass the online face comparison service. Compared with BMI-S, BMI is less robust since it has a larger search space than BMI-S, and thus is easier to get stuck in local minima. Meanwhile, FD fails to provide meaningful results because of the limitation of its computational complexity.

4.4. Method Analysis

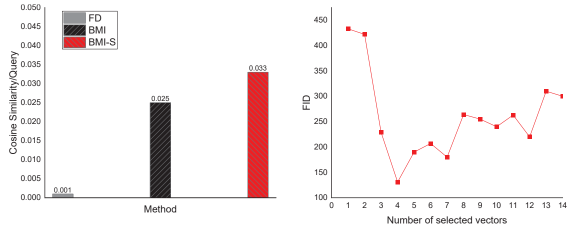

Fig. 3. Comparison of gradient estimation accuracy of different methods and FID with different number of key style vectors.

For the analysis of BMI-S, we do some validation experiments to evaluate the effects of (1) NES gradient estimation and (2) key style vector selection. For the comparison of gradient estimation, we calculate the cosine similarity of estimated gradients produced by three different methods and true gradients (white-box gradients), then divide the cosine similarity with the query numbers required to get the results. According to Fig 3, compared with FD, BMI and BMI-S significantly increase the estimation accuracy per query. Meanwhile, the mean FID score varies with the variation of the number n that style control vectors are selected, where too small n can not provide sufficient variation while too large n may lead to larger search space and less robustness. According to Fig 3, $n = 4$ is the best choice under this condition.

5. CONCLUSION

In this paper, we focus on face recognition tasks and explore model inversion attacks under black-box constraints. We first prove that training surrogate models for deployed face recognition systems is difficult. Then we propose BMI-S, a sparse black-box model inversion attack. BMI-S uses Natural Evolution Strategy to perform efficient gradient estimation, which enables query-based attacks. In addition, we accelerate the gradient estimation process by selecting and only optimizing the key style vectors of the image generator. Experimental results on widely used datasets show that BMI-S can achieve high-quality model inversion attacks with limited information and queries.

6. REFERENCES

[1] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon M. Lin, David Page, and Thomas Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *USENIX Security Symposium*. 2014, pp. 17–32, USENIX Association.

[2] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *CCS*. 2015, pp. 1322–1333, ACM.

[3] 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 *CVPR*. 2020, pp. 250–258, Computer Vision Foundation / IEEE.

[4] Shagufta Mehnaz, Sayanton V. Dibbo, Ehsanul Kabir, Ninghui Li, and Elisa Bertino, "Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models," in *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, Kevin R. B. Butler and Kurt Thomas, Eds. 2022, pp. 4579–4596, USENIX Association.

[5] Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correia, Antonia Adler, and Kristian Kersting, "Plug & play attacks: Towards robust and flexible model inversion attacks," in *ICML*. 2022, vol. 162 of *Proceedings of Machine Learning Research*, pp. 20522–20545, PMLR.

[6] Mahdi Khosravy, Kazuaki Nakamura, Yuki Hirose, Naoko Nitta, and Noboru Babaguchi, "Model inversion attack by integration of deep generative models: Privacy-sensitive face generation from a face recognition system," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 357–372, 2022.

[7] Shagufta Mehnaz, Ninghui Li, and Elisa Bertino, "Black-box model inversion attribute inference attacks on classification models," *CoRR*, vol. abs/2012.03404, 2020.

[8] Ulrich Aïvodji, Sébastien Gambs, and Timon Ther, "GAMIN: an adversarial approach to black-box model inversion," *CoRR*, vol. abs/1909.11835, 2019.

[9] Thomas Bekman, Masoumeh Abolfathi, Jafar Haadi Jafarian, Ashis Biswas, Farnoush Banaei Kashani, and Kuntal Das, "Practical black box model inversion attacks against neural nets," in *PKDD/ECML Workshops (2)*. 2021, vol. 1525 of *Communications in Computer and Information Science*, pp. 39–54, Springer.

[10] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*. PMLR, 2016, pp. 1558–1566.

[11] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin, "Black-box adversarial attacks with limited queries and information," in *ICML*. 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 2142–2151, PMLR.

[12] Andrew Ilyas, Logan Engstrom, and Aleksander Madry, "Prior convictions: Black-box adversarial attacks with bandits and priors," in *ICLR (Poster)*. 2019, OpenReview.net.

[13] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang, "Heuristic black-box adversarial attacks on video recognition models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 12338–12345.

[14] Tobias Glasmachers, Tom Schaul, and Jürgen Schmidhuber, "A natural evolution strategy for multi-objective optimization," in *PPSN (1)*. 2010, vol. 6238 of *Lecture Notes in Computer Science*, pp. 627–636, Springer.

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," in *CVPR*. 2020, pp. 8107–8116, Computer Vision Foundation / IEEE.

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

[17] S. Milborrow, J. Morkel, and F. Nicolls, "The MUCT Landmarked Face Database," *Pattern Recognition Association of South Africa*, 2010, http://www.milbo.org/muct.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*. 2016, pp. 770–778, IEEE Computer Society.

[19] Alexander Mathiasen and Frederik Hvilshøj, "Fast fréchet inception distance," *CoRR*, vol. abs/2009.14075, 2020.

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.