Poster: Query-Efficient Adversarial Attack Framework by Utilizing Transfer Gradient

Chilin Fu
Ant Financial Services Group
chilin.fcl@antfin.com

Xiaolu Zhang
Ant Financial Services Group
yueyin.zxl@antfin.com

Jun Zhou
Ant Financial Services Group
jun.zhoujun@antfin.com

***Abstract*—In this work, we propose a heuristic framework, which combines the transfer-based method with the query-based method to improve the black-box adversarial attack. The framework is composed of the transfer part and the adaptive query part, which can maximize the benefit of transfer-based prior and improve query performance. The skip query strategy is adopted to accelerate the gradient estimation procedure, and the transfer gradient is integrated into query information to avoid the convergence problem. Experiments consistently demonstrate that compared with the baseline methods, our methods require much fewer queries to perform black-box attacks. Moreover, our framework is flexible to integrate other query-efficient methods such as dimension reduction techniques to further improve query efficiency.**

I. INTRODUCTION

Though Deep Neural Networks(DNNs) have achieved significant success in various applications, recent research has demonstrated that even state-of-the-art neural networks are vulnerable to adversarial examples [1]. In this work, we consider the black-box adversarial setting, which assumes that the attackers have no or limited access to the target model and are more practical in real-world situations. Based on whether an attacker needs to query the target model, black-box attacks can be roughly divided into transfer-based methods and query-based methods. However, query-based methods usually require a tremendous number of queries to estimate the approximate gradient [2], and transfer-based methods have no guarantee of a satisfactory attack success rate [3], [4].

Recently, some work has combined these methods to improve query efficiency [5], [6]. The Prior-RGF integrates the gradient of a surrogate model into the gradient estimated by the query information through an optimal coefficient that controls the strength of the transfer gradient. Following the idea of using a transfer gradient as prior information, we propose a heuristic framework, which always maximizes the benefit of the transfer gradient and improves the performance. The framework can be divided into two parts, *the transfer part* and *the adaptive query part*. When the transferability of the surrogate model is strong enough, or the adversary is not close to the adversarial region, we only utilize the transfer gradient to update the adversary. Besides, while the gradient of the surrogate model points to the non-adversarial region we use an early stop strategy to terminate this procedure. In the adaptive-query part, since successive gradients are heavily correlated along the gradient estimation trajectory [7], we adopt the strategy of skipping queries and reusing the gradient estimated in previous iterations. Furthermore, to accelerate query efficiency and avoid the convergence problem caused by the skipping strategy at the beginning stage, we integrate the transfer gradient and query gradient as the overall estimation gradient.

We evaluate our method on the ImageNet validation dataset by comparing it with the alternative state-of-the-art methods. The results demonstrate that our method requires much fewer queries than the baseline method without sacrificing the attack success rate. Furthermore, the proposed framework is flexible to integrate various query-efficient methods such as those dimension reduction techniques or other prior information.

II. PROBLEM STATEMENT

We consider the target model that the class prediction scores are known to an attacker. Following this setting, we denote the black-box model as a classification function $f(x) \in \mathbb{R}^K$, $x \in [0,1]^D$ is the input image with D dimensions, K is the number of classes, and the model yields a vector of prediction probabilities of all K image classes. The cross-entropy loss function can be denoted as $J(f(x), y)$, which depends on the output $f(\cdot)$ and the desired class label y. The goal of attack is to generate an adversarial example x^{adv} that is classified as target label while the L_p norm of the adversarial noise is less than an allowed value ϵ as

$$\arg\max_k f(x^{adv}) = y, \ \ s.t. ||x^{adv} - x||_p \leq \epsilon. \qquad (1)$$

III. METHODOLOGY

In this section, we illustrate the details of our framework, which includes the transfer part and the adaptive query part. We first use the transfer part to craft an input example, and the adaptive query part will be executed if the output of transfer part is not a successful adversarial example. The projected gradient descent(PGD) [8] is iteratively used to generate the adversarial example in both parts, except that the true gradient is replaced by an approximate gradient.

The Transfer Part. As mentioned above, the transfer part only utilizes the gradient that is obtained by the surrogate model to update the input example. An early stop strategy is proposed to prevent the adversary to move towards the non-adversarial region. Specifically, we query the black-box model

TABLE I
THE EXPERIMENTAL RESULTS OF BLACK-BOX ATTACKS

Methods	Inception-v3		VGG-16		ResNet-50	
	ASR	AVG.Q	ASR	AVG.Q	ASR	AVG.Q
Prior-RGF(ResNet-152)	92.2%	1489	98.2%	1298	98%	1197
Prior-RGF(MobileNet-V2)	92%	1525	98.3%	1319	97.6%	1634
ours(ResNet-152,S=0)	**95.1%**	1177	**99.7%**	833	**99.9%**	569
ours(ResNet-152,S=1)	92.5%	**794**	98.5%	**573**	99.2%	**431**
ours(MobileNet-V2,S=0)	**94.2%**	1256	**99.7%**	895	**99.1%**	1065
ours(MobileNet-V2,S=1)	91.5%	**851**	98.5%	**627**	97.5%	**844**

to compute the value of loss function J at the end of each step and terminate the iteration if the loss value does not decrease in N consecutive steps, which is set as 10 in our experiments.

The Adaptive Query Part. According to [7] successive gradients are heavily correlated, we skip queries by reusing the gradient estimated in the previous interations. The number of steps can skip is denoted as S, and we take $S = 0$ and $S = 1$ in the following experiment. The basic gradient estimation method we adopt is the Random Gradient-Free(RGF) [9], in each query step, the gradient is estimated by

$$g = \frac{1}{q}\sum_{i=1}^{q} g_i, \;\; g_i = \frac{J(f(x+\sigma u_i), y) - J(f(x), y)}{\sigma} \cdot u_i \quad (2)$$

where u_i is a random Gaussian vector, and $\sigma > 0$ is a smoothing parameter. g is the average estimation over q random directions to reduce the variance. In addition, we combine the transfer gradient v with the estimated gradient g to obtain the approximate gradient $\hat{g}$ as

$$\hat{g} = w \cdot g + (1 - w) \cdot v \quad (3)$$

where w is the coefficient to balance the estimated gradient and the transfer gradient. The purpose of using the transfer gradient here is to alleviate the convergence problem mentioned above and accelerate query efficiency. The weight w is increasing with the iteration until it is greater than or equal to 1.

IV. EXPERIMENT SETUP

We evaluate our approach for attacking black-box models in targeted attacks setting under l_∞ norm on the ImageNet validation dataset. We randomly select 1,000 images that are correctly predicted by all black-box models and fixed a random target label for each image. We use the ResNet-152 model and MobileNet-V2 model as the surrogate model to generate transfer gradient, and attack the three normal training models of Inception-V3, VGG-16, and ResNet-50. All of the above models are provided by torchvision for ImageNet. We set the perturbation size as $\epsilon = 0.05$, learning rate as $\eta = 0.01$ for PGD under the l_∞ norm, with images in $[0, 1]^D$, $D = 224 \times 224 \times 3$. We restrict the maximum number of queries $Q = 10,000$ for each attack, if the adversarial example is generated within the Q queries, it can be considered a successful attack. We compare the proposed method with the baseline method Prior-RGF with the derived optimal λ^*.

The experimental results are shown in Table 1. In these methods, we set the number of query used to estimate the gradient to $q = 50$, and the sampling variance to $\sigma = 0.0001$. The initial weight of the gradient estimated in the skip step is set to $w = 0.5$, and each skip step is multiplied by 1.1, until it reaches 1.0. We use the *attack success rate* and the *average number of query* as metrics for comparing with the baseline method. It can be seen that our method needs much fewer query than the baseline without sacrificing the attack success rate. Besides, we limit the maximum number of query for each adversarial example and analyze the success rate over these methods, the results are shown in Fig. 1, which demonstrates that our method has higher success rate than the baseline method with fewer query budgets.

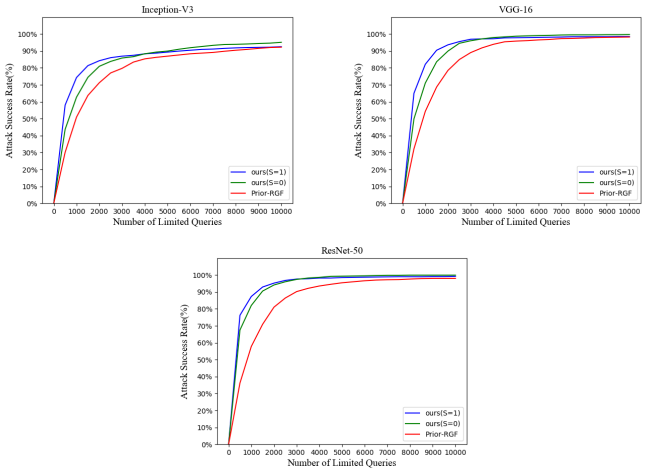

Fig. 1. Attack Success Rate with different query budgets.

V. CONCLUSION

In this paper, we propose a framework that can always maximize the benefit of the transfer gradient for improving the query efficiency of black-box attack. The experimental results demonstrate the effectiveness of our method, which requires much fewer queries compares to the baseline method. Furthermore, our framework is flexible to integrate various query-efficient methods such as those dimension reduction techniques or other prior information.

REFERENCES

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2014.

[2] Y. Liu, X. Chen, C. Liu, and D. X. Song, "Delving into transferable adversarial examples and black-box attacks," in *ICLR*, 2016.

[3] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *ICLR*, 2017.

[4] H. Li, X. Xu, X. Zhang, and S. Yang, "Qeba: Query-efficient boundary-based blackbox attack," in *CVPR*, 2020.

[5] S. heng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving black-box adversarial attacks with a transfer-based prior," in *NeurIPS*, 2019.

[6] J. Du, H. Zhang, J. T. Zhou, Y. Yang, and J. Feng, "Query-efficient meta attack to deep neural networks," in *ICLR*, 2020.

[7] A. Ilyas, L. Engstrom, and A. Madry, "Prior convictions: Black-box adversarial attacks with bandits and priors," in *ICLR*, 2019.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2017.

[9] Y. Nesterov and V. G. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, pp. 527–566, 2017.