# Abstract

With the development of adversarial machine learning, there is an increasing demand for effective and efficient black-box adversarial attacks. Transfer-based and query-based attacks have achieved high attack success rates in various scenarios. However, these two types of black-box attacks suffer from a lack of effectiveness or efficiency. Recently, novel attacks combining these two types of attacks have achieved state-of-the-art performance, demonstrating the potential of this new type of attack.

Through a deep study on transfer-based black-box adversarial attacks, find that gradients computed on alternative white-box models have different prior knowledge about the target model with gradients estimated by queries on the target black-box model. Motivated by this, a transfer-based black-box attack, NI-FGSM (Nesterov Iterative Fast Gradient Sign Method), is improved by introducing the RGF (Random Gradient-Free) estimated gradient of the target black-box model. Propose two novel black-box attacks combining transfer-based and query-based attacks. (1) LAQ-NI-FGSM (Look Ahead towards Query gradient NI-FGSM) changes the forward-looking direction in NI-FGSM to the weighted average of the gradient estimated by queries and the original forward-looking direction. (2) Based on LAQ-NI-FGSM, 2NI-FGSM (2-gradients-guided NI-FGSM) then applies the cumulative momentum method to the query gradients. Apart from that, 2NI-FGSM changes the updating direction of adversarial examples to the weighted average of the query momentum and the original updating direction.

Extensive experiments demonstrate the effectiveness and efficiency of the purposed 2NI-FGSM. The experimental results show that the attack success rates of 2NI-FGSM on multiple undefended models exceed those of the previous attacks by 7-15 percentage points, and the average number of queries is 15-20% lower.

**Keywords:** Adversarial examples, Black-box attacks, Transferability, Gradient estimation