# Improving Query Efficiency of Black-box Adversarial Attack

Yang Bai [1] Yong Jiang [1]

[1]Tsinghua Berkeley Shenzhen Institute, Tsinghua University, accepted by ECCV2020.

## Abstract

Deep neural networks (DNNs) are under the risk of adversarial examples that can be easily generated when the target model is accessible to an attacker (white-box setting). As plenty of machine learning models have been deployed via online services that only provide query outputs from inaccessible models (*e.g.*, Google Cloud Vision API2), black-box adversarial attacks (inaccessible target model) are of critical security concerns in practice rather than white-box ones. However, existing query-based black-box adversarial attacks often require excessive model queries to maintain a high attack success rate. Therefore, in order to improve query efficiency, we propose a Neural Process based black-box adversarial attack (NP-Attack). Extensive experiments show that NP-Attack could greatly decrease the query counts under the black-box setting.

## Introduction

Deep neural networks (DNNs) have been deployed on many real-world complex tasks and demonstrated excellent performance. However, DNNs are found vulnerable to adversarial examples crafted following either a white-box setting (the adversary has full access to the target model) or a black-box setting (the adversary has no information of the target model). Among the black-box ones, *query*-based black-box attacks that directly generate adversarial examples on target models are proposed. These query-based methods could bring almost 100% attack success rates however with very high query counts (not acceptable). Therefore, how to significantly reduce the query complexity while maintaining the attack success rate simultaneously is still an open problem.

In this paper, we introduce the structure information of the image into consideration to further reduce the query counts. To be specific, the structure is characterized by a Neural Process (NP), an efficient auto-encoder method to model a distribution over regression functions.

Our main contributions could be summarized as follows:

- We propose a distribution based black-box attack, Neural Process based black-box attack (NP-Attack), which uses the image structure information for modeling adversarial distributions and can reduce the required query counts.
- Extensive experiments demonstrate the superiority of our proposed NP-Attack on both untargeted and targeted attacks.

## NP-Attack

Our method NP-Attack mainly includes two procedures, the pre-training of NP model and the main-stream black-box attack.

## Algorithm

**Algorithm 2** NP-Attack

**Input:** image data $x$, label $y$ (if untargeted attack $y$ represents the true label of $x$; else $y$ represents the target label), target neural network $F$, pre-trained NP model (*Encoder* $h$, *Decoder* $g$, *Aggregator* $a$), maximal optimization iteration $T$, sample size $b$, projecting function $P$, learning rate $\eta$

Compute the variables from *Encoder* $h$ on image $x$: $\mathcal{N}(\mu, \sigma^2), r \leftarrow h(x)$

**for** $t = 0$ **to** $T - 1$ **do**

  **if** F(x) $\neq$ y (untargeted) or F(x) = y (targeted) **then**

    attack success.

  **else**

    Sample $b$ independent random perturbations from standard Gaussian distribution: $p_i \sim \mathcal{N}(0, I), i = 1, 2, ..., b$.

    Add the perturbation $p_i$ in NP-Attack-R/Z/RZ, specifically on $\mu$ or $r$ or both,

$$\begin{cases} z_i \sim \mathcal{N}(\mu, \sigma^2), r_i = r + p_i\sigma, & \text{NP-Attack-R,} \\ z_i \sim \mathcal{N}(\mu + p_i\sigma, \sigma^2), r_i = r, & \text{NP-Attack-Z,} \\ z_i \sim \mathcal{N}(\mu + p_i\sigma, \sigma^2), r_i = r + p_i\sigma, & \text{NP-Attack-RZ.} \end{cases}$$

    Reconstruct the image from *Decoder* $g$, and use project function $P$ to restrict its maximal distortion: $x_i = P(g(z_i, r_i))$.

    Compute the losses of these reconstructed image series $x_i$ under targeted or untargeted setting,

$$L_i = \begin{cases} \max(0, \max_{c \neq y} \log F(x_i)_c - \log F(x_i)_y), & \text{targeted,} \\ \max(0, \log F(x_i)_y - \max_{c \neq y} \log F(x_i)_c), & \text{untargeted.} \end{cases}$$

    The corresponding loss $L_i = L_i - mean(L)$.

    Update $\mu$ or $r$ or both as optimization:

$$\begin{cases} r_{t+1} \leftarrow r_t - \frac{\eta}{b\sigma} \sum_{i=1}^{b} L_i p_i, & \text{NP-Attack-R,} \\ \mu_{t+1} \leftarrow \mu_t - \frac{\eta}{b\sigma} \sum_{i=1}^{b} L_i p_i, & \text{NP-Attack-Z,} \\ \mu_{t+1} \leftarrow \mu_t - \frac{\eta}{b\sigma} \sum_{i=1}^{b} L_i p_i, r_{t+1} \leftarrow r_t - \frac{\eta}{b\sigma} \sum_{i=1}^{b} L_i p_i, & \text{NP-Attack-RZ.} \end{cases}$$

  **end if**

**end for**

## Experiments

**CIFAR-10** The black-box attack results on CIFAR-10.

Table 1:Adversarial evaluation of black-box attacks on CIFAR10.

| Attack Method | Untargeted Attack | | | | Targeted Attack | | | |
|---|---|---|---|---|---|---|---|---|
| | ASR | $L_2$ | $L_\infty$ | Query Count | ASR | $L_2$ | $L_\infty$ | Query Count |
| ZOO | 100% | 0.12 | 0.02 | 208,384 | 99.52% | 0.19 | 0.02 | 230,912 |
| AutoZOOM-BiLIN | 100% | 1.56 | 0.15 | 8,113 | 100% | 2.13 | 0.21 | 8,266 |
| AutoZOOM-AE | 100% | 1.88 | 0.16 | 7,113 | 100% | 2.78 | 0.24 | 8,217 |
| QL | 98.40% | 1.91 | 0.05 | 857 | 99.55% | 2.11 | 0.05 | 616 |
| $\mathcal{N}$Attack | 99.89% | 2.61 | 0.05 | 183 | 100% | 2.61 | 0.05 | 1,151 |
| NP-Attack-R(Ours) | 100% | 1.74 | 0.05 | **94** | 100% | 1.85 | 0.05 | **589** |
| NP-Attack-Z(Ours) | 100% | 1.67 | 0.05 | 144 | 100% | 1.78 | 0.05 | 936 |

**ImageNet** The black-box attack results on ImageNet.

Table 2:Adversarial evaluation of black-box attacks on ImageNet.

| Attack Method | Untargeted Attack | | | Targeted Attack | | |
|---|---|---|---|---|---|---|
| | ASR | $L_2$ Dist | Query Count | ASR | $L_2$ Dist | Query Count |
| ZOO | 90% | 1.20 | 15,631 | 78% | 3.43 | $2.11 \times 10^6$ |
| AutoZOOM-BiLIN | 100% | 9.34 | 3,024 | 100% | 11.26 | 14,228 |
| QL | 100% | 17.72 | 3,985 | 100% | 17.39 | 33,360 |
| $\mathcal{N}$Attack | 100% | 24.01 | 2,075 | 100% | 24.14 | 14,229 |
| Bandits | 100% | -- | 3,165 | 92.5% | -- | 25,341 |
| NP-Attack-R(Ours) | 100% | 10.96 | **867** | 98.02% | 14.38 | **8,001** |
| NP-Attack-Z(Ours) | 96.04% | 12.37 | 1,236 | 98.02% | 14.60 | 11,383 |

The visual results of our NP-Attack on ImageNet. The perturbations added by our NP-Attack are imperceptible.

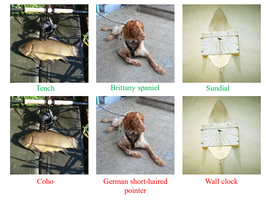


Figure 1:Adversarial example of ImageNet generated by NP-Attack-R on untargeted attack bounded by $L_\infty$=0.05. Top row shows the original images and the true labels (marked as green) while the bottom row are the adversarial examples with predicted label (marked as red). The visual results perform similar on MNIST and CIFAR-10.

## References

[1] Marta Garnelo, Jonathan Schwarz, Rosenbaum Dan, Fabio Viola, and Yee Whye Teh. Neural processes. In *ICLR*, 2018.

[2] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018.

[3] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*, 2018.

[4] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In *ICML*, 2019.

[5] Xinyi Liu, Yang Bai, Shu-Tao Xia, and Yong Jiang. Self-adaptive feature fool. In *ICASSP*, pages 4177--4181. IEEE, 2020.