CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2022-0005

Vol. **4**, No. 4, pp. 797-819
December 2023

Non-Lipschitz Attack: A More Sparse Adversarial Attack via Non-Lipshitz ℓ_p Regularization

Xuan Lin[1,2], Haidong Xie[3,*], Chunlin Wu[2] and Xueshuang Xiang[1]

[1] *Qian Xuesen Laboratory, China Academy of Space Technology, Beijing, P.R. China.*
[2] *School of Mathematical Sciences, Nankai University, Tianjin, P.R. China.*
[3] *China Academy of Aerospace Science and Innovation, Beijing, P.R. China.*

Received 20 January 2022; Accepted 20 July 2023

Abstract. Deep neural networks are considerably vulnerable to adversarial attacks. Therein, sparse attacks mislead image classifiers with a sparse, pixel-level perturbation that alters few pixels, and have much potential in physical world applications. The existing sparse attacks are mostly based on ℓ_0 optimization, and there are few theoretical results in these works. In this paper, we propose a novel sparse attack approach named the non-Lipschitz attack (NLA). For the proposed ℓ_p $(0<p<1)$ regularization attack model, we derive a lower bound theory that indicates a support inclusion analysis. Based on these discussions, we naturally extend previous works to present an iterative algorithm with support shrinking and thresholding strategies, as well as an efficient ADMM inner solver. Experiments show that our NLA method outperforms comparative attacks on several datasets with different networks in both targeted and untargeted scenarios. Our NLA achieves the 100% attack success rate in almost all cases, and the pixels perturbed are roughly 14% fewer than the recent ℓ_0 attack FMN-ℓ_0 on average.

AMS subject classifications: 68T07

Key words: Sparse adversarial attack, ℓ_p $(0<p<1)$ regularization, lower bound theory, support shrinkage, ADMM.

1 Introduction

Recent studies have illustrated the vulnerability of deep neural networks (DNNs) to adversarial examples (AEs) [35,63] that are custom-designed to mislead classifications. The existing adversarial attacks designed to craft AEs have achieved the 100% attack success rate (ASR) [3, 6, 10]. It is natural to have the following ideas to generate AEs similar to the original samples [6, 10, 48, 63] to reduce the consumption: we either generate a small

*Corresponding author. *Email addresses:* 1120200026@mail.nankai.edu.cn (X. Lin), xiehaidong@aliyun.com (H. Xie), wucl@nankai.edu.cn (C. Wu), xiangxueshuang@qxslab.cn (X. Xiang)

perturbation on each pixel or craft a sparse perturbation that only on a few pixels. To facilitate the usage of those attacks in the physical world, we are more concerned about yielding sparse perturbation.

In sparse optimization theory, there are approximately three kinds of methods, named ℓ_0 [25], ℓ_1 [11,21,64], and ℓ_p ($0<p<1$) regularized [12,55] optimization, where ℓ_1-optimization is a convex problem and the other two are nonconvex. In general, the solution to the nonconvex model is sparser than that to the convex ℓ_1 model. It has been shown that the solution obtained from ℓ_p minimization methods are much more sparse than that from ℓ_1 methods [8, 9, 31, 68, 76] in compressive sensing. To our knowledge, ℓ_1 attacks include B&B-ℓ_1, ℓ_1-APGD [20], and related elastic-net attack (EAD) [10]. Since ℓ_0 minimization is NP-hard, different ways have been built to approximate the ℓ_0 minimization to construct attack methods, such as Jacobian-based saliency map attack (JSMA) [53] and its variants [16, 65], C&W-ℓ_0 attack [6], one-pixel attack [62] and its variant [38], adversarial patch (AdvPatch) [4], B&B-ℓ_0 attack [3], and FMN-ℓ_0 attack [54]. These works have excellent performances with few theoretical analyses. At present, we have not found any reference discussing the combination of adversarial attacks and ℓ_p ($0<p<1$) regularization.

Actually, ℓ_p minimization has recently shown remarkable performance in signal and image restoration [13, 30, 52, 66, 71], especially on the recovery of edges. It is also applied in the objective functions of different DNNs [1, 46]. Nevertheless, the ℓ_p minimization remains a challenging problem since it gives rise to a nonconvex nonsmooth non-Lipschitz optimization problem. Most existing approaches to non-Lipschitz optimization are approximation methods that introduce auxiliary parameters to transform the non-Lipschitz models into Lipschitz models.

To obtain sparser perturbations, a non-Lipschitz ℓ_p regularized optimization framework, named the non-Lipschitz attack (NLA), is proposed to try to craft sparser AEs. For our ℓ_p regularization model, we first establish a lower bound theory for all limiting stationary points (including local minimizers, even global ones) of the objective function of the model for the adversarial attack. By a support inclusion motivation analysis, we extend previous works to design an iterative implementation using support shrinking and thresholding strategies. In addition, we adopt the alternating direction method of multipliers (ADMM) to efficiently solve a subproblem in each inner loop. As we expect to obtain the sparse perturbation with ASR nearing 100%, the trade-off binary search technique and the threshold operation (resulting from the lower bound theory) for the variable in each iteration are adopted to achieve these goals, that is, to select the parameters adaptively to craft a sparse perturbation with ASR 100% as possible. Fig. 1 visually displays the technological flow process of NLA, as well as the sparsity and adversary of a given sample.

We compare our NLA to several recent attacks on datasets MNIST and CIFAR10 with several ordinary networks, showing that NLA achieves a 100% attack success rate in all cases and reduces the ℓ_0 distortion by roughly 14% on average compared to the leading level sparse attack FMN-ℓ_0.