Copyright ©2022 The Institute of Electronics,
Information and Communication Engineers

SCIS 2022 2022 Symposium on
Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics,
Information and Communication Engineers

ベイズ最適化を用いたデータ・クエリ効率の良い Black-box Universal Adversarial Attacks

Data and Query Efficient Black-box Universal Adversarial Attacks with Bayesian Optimization

由比藤 真 * 米山 一樹 *
Makoto Yuito Kazuki Yoneyama

キーワード Black-box Universal Adversarial Attacks, ベイズ最適化, AI セキュリティ

あらまし

Adversarial Examples (AE) による Adversarial Attacks は Deep Neural Network (DNN) における最大の脆弱性のひとつであり、White-box 及び Black-box 攻撃の研究が盛んに行われている。最近では、Universal Adversarial Perturbation (UAP) と呼ばれる、任意の画像に加えることで AE を生成することができる単一の摂動を計算する Universal Adversarial Attacks (UAA) が研究されている。

UAA の既存研究は主に White-box 環境を想定しているものが多い。いくつかの既存研究は、クエリアクセスのみを用いる Black-box 環境下で UAP を生成できることを示しているが、その多くは Black-box 環境で高い攻撃成功率を達成するために、大量の訓練データとターゲットモデルへのクエリ回数を必要としており、実際の MLaaS 等へ適用することを考えると現実的ではないという問題がある。

本稿では、より現実的なセッティングに基づく Black-box UAA を考える。具体的には、ごく少量の訓練データと最低限のクエリ回数のみで UAP を生成することを目標とする。効率よく UAP を生成するために、我々はベイズ最適化を用いた新しい Black-box UAA 手法を提案する。

ImageNet を分類する 3 つのモデルにおいて、図 1 に示す訓練データ数とクエリ回数を用いて UAP を生成し、訓練データとは独立するテストデータにおける提案手法と既存手法の攻撃成功率を比較する。提案手法は既存手法よりも極めて少ない訓練データ数 (50 クラス分) とクエリ回数を用いて同等の攻撃成功率 (最高 81%) を達成する。この結果は、提案手法によって生成された UAP が高いクラス一般化性能を持ち、1000 クラスの画像で構成されるテストデータの多くを誤分類させることを可能にすることを示している。

また、より多くの AE を生成することを目的に、画像単位で AE を生成する Adversarial Attacks とクエリ効率の観点から比較を行う。結果として、一定の割合を超える AE を生成したい場合に、提案手法が最先端の Adversarial Attacks 手法のクエリ効率を上回ることを示す。

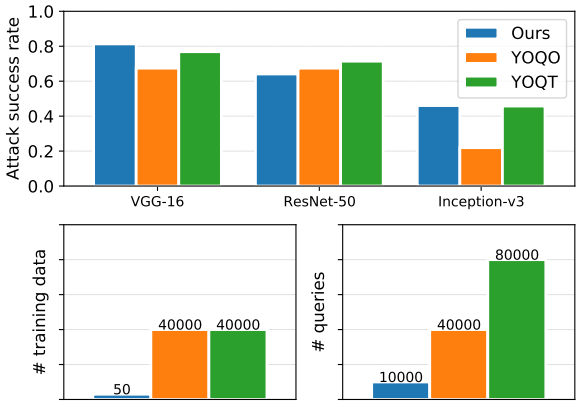

図 1: ImageNet を分類する VGG-16、ResNet-50、Inception-v3 アーキテクチャのモデルにおける、既存手法及び提案手法の攻撃成功率、訓練データ数、クエリ数の比較。

* 茨城大学, 茨城県日立市中成沢町 4 丁目 12-1, Ibaraki University, 4-12-1, Nakanarusawa-cyo, Hitachi-city, Ibaraki, 316-8511 Japan. {20nm733a,kazuki.yoneyama.sec}@vc.ibaraki.ac.jp