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

シンプルブラックボックス攻撃の対策手法に関する検討
Designing Countermeasures of the Simple Black-box Attack

鵜島 廉 * 井林 大成 * 満保 雅浩 *
Ren Ujima Taisei Ibayashi Masahiro Mambo

キーワード Adversarial Example, 深層学習

あらまし

近年、Deep Neural Network(DNN) の精度が向上し、画像処理など様々な分野で利用されるようになっている。機械学習モデルに誤分類をさせる攻撃の一つに、Adversarial Example が存在する [1]。この攻撃は、DNN モデルに入力する画像に摂動を加えることで、誤分類を引き起こす入力を生成する攻撃である。

Adversarial Example は、攻撃者がモデルの内部パラメータを利用できる White-Box 攻撃と、情報が制限される Black-Box 攻撃の 2 つに分類できる。攻撃対象のモデルがインターネット上などにある場合では、パラメータを利用できない場合がほとんどであるため、Black-Box 攻撃への対策が重要である。

本論文ではモデルの出力確率の変化を利用する単純な Black-Box 攻撃である、Simple BlackBox Attack の対策手法を提案、実装、評価を行った。一般的に推論モデル内に分類器は 1 つであるが、提案手法では複数の分類器を使用する。推論モデルへの入力に対して、ランダムに分類器を選択し、その推論結果をモデルの出力とする。分類器ごとに出力確率には差があることから、攻撃者が正しい摂動を選択できる可能性を下げ、攻撃にかかる時間及びモデルへのアクセス回数の増加や、Adversarial Example の画質を悪くする効果が期待できる。

提案手法の評価実験として、通常の分類器が 1 つのモデルと、複数のモデルの場合で比較実験を行う。データセットは MNIST と CIFAR-10 を使用する。提案手法では分類器間の出力の違いを利用するため、訓練データセットを 2 等分しそれぞれ訓練を行うことが考えられる。評価基準として、画像の劣化を示す指標である PSNR と、攻撃を繰り返した回数であるクエリ数を使用する。これにより攻撃を終えるまでの時間変化や、変化した画像が人間によって検知しやすくなったかどうかを評価することができる。評価実験の 1 つでは、攻撃者が 1 度に加える摂動の大きさを変化させ、そのときのクエリ数などの変化を観察する。画素が 256 階調で表現されているため、一度に乗せる摂動の量を 1/255,2/255,$\cdots$ のように変化させる。結果として、まずクエリ数は摂動の変化量が小さいときほど、提案手法で高くなり、摂動の変化量が大きくなると通常のモデルとの差が小さくなり、値がほぼ一定となることを示す。PSNR は、摂動の変化量が大きいほど、通常のモデルとの差は小さくなるが、提案手法では PSNR が通常のモデルより低下する傾向となる。

このほかの実験の結果も含めて、提案手法が Simple Black Box Attack による摂動の選択を難化させる効果があることを示す。

参考文献

[1] Christian Szegedy,Wojciech Zaremba,Ilya Sutskever,Joan Bruna,Dumitru Erhan, Ian Goodfellow,and Rob Fergus, ”Intriguing properties of neural networks ”,arXiv:1312.6199, 2013

* 金沢大学, 〒 920-1192 石川県金沢市角間町, Kanazawa University,Kakuma-machi, Kanazawa-shi, Ishikawa 920-1192, Japan