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

JPEG 圧縮由来の歪み信号に対する応答特性に基づく Adversarial Examples 検知手法

Detection of Adversarial Examples Using Sensitivities to JPEG Compression-Derived Signals

角森 健太 * 山﨑 裕真 * 栗林 稔 * 舩曳 信生 *
Kenta Tsunomori Yuma Yamasaki Minoru Kuribayashi Nobuo Funabiki

越前 功 †
Isao Echizen

キーワード Adversarial Examples，畳み込みニューラルネットワーク，JPEG 圧縮，スケーリング

あらまし

画像分類や音声認識などでは深層学習の技術が活用されている．深層学習の代表的な手法の一つである畳み込みニューラルネットワーク (CNN) を用いた画像分類器は，高い分類精度を示している．一方で，入力画像に対して微小な敵対的ノイズを加えて作成される Adversarial Examples(AEs) は画像分類器の誤分類を引き起こすことが報告されている．このノイズは人間の目視による確認が難しい．そのため，AEs に対応する技術開発は，画像分類器を利用する上で必要不可欠である．

AEs に対する防御手法の一つとして，入力画像が AEs であるか事前に検知する方法が挙げられる．東ら [1] は入力画像に対して強度の異なるノイズ除去フィルタを使用した際の画像分類器の応答特性に着目した手法を提案した．JPEG 圧縮やスケーリングは，フィルタ強度が大きくなるほど画像の情報量を減らし，画像に含まれているノイズを除去すると考えられる．画像のノイズ除去は，画像分類器の出力に変化を与えるため，東らはこの出力の変化を特徴と考え，AEs の検知システムに用いていた．この手法では 14 種ものフィルタを使用しており，計算量が多いことが問題となっていた．

本稿ではフィルタ数を減らし，かつ高い精度で AEs を検知する手法を提案する．提案手法では，先行手法に JPEG 圧縮前と圧縮後の画像の差分を利用したノイズ除去フィルタを加える．東らの報告から，AEs に対して JPEG 圧縮を適用すると敵対的ノイズがうまく除去されると考えられる．このことから，JPEG 圧縮前と圧縮後の画像の差分は加えられた敵対的ノイズと高い相関を有すると予想される．そこで本研究では，この差分の大きさ・正負を変化させて画像に加え，敵対的ノイズを効果的に除去するアプローチを取る．また，この差分内の各要素の正負を反転させる割合を変化させることでノイズ除去精度も変化し，画像分類器の出力にも影響を与えると考えられる．更には，このように変化させた差分を画像に加えた後にスケーリング処理を施すことによって，取り除き切れなかったノイズまでうまく除去することを試みる．

シミュレーションでは，非標的型攻撃と標的型攻撃で Adversarial Examples 検知精度の評価を行った．非標的型攻撃において提案手法は，ほとんどの攻撃で先行手法より検知精度が高くなり．標的型攻撃においては，先行手法に対する精度低下を 0.5%以内に抑えた．

参考文献

[1] 東 亮憲，栗林 稔，舩曳 信生，Huy Hong Nguyen，越前 功，“複数のフィルタ強度による CNN 画像分類器の応答特性を用いた敵対的事例の検出法，” 信学技報，vol.120，no.418，EMM2020-70，pp.19-24，2021 年

* 岡山大学，〒 700-8530 岡山県岡山市北区津島中三丁目 1 番 1 号, Okayama University, 3-1-1 Tsushimanaka, Kita, Okayama, Okayama
† NII, 〒 101-8430 東京都千代田区一ツ橋 2-1-2, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo