博士請求論文審査要旨

情報セキュリティ大学院大学
情報セキュリティ研究科

論文題目 ： Automated Detection System for Adversarial Examples on Images with Image Transformation and Filters

申請者 ： Dang Duy Thang

審査委員会 ： 主査　教授　後藤 厚宏
副査　教授　湯淺 墾道
副査　教授　大塚 玲
副査　教授　松井 俊浩

Ⅰ. 論文内容の要旨

AI において，深層学習ニューラルネットワーク(以下 DLNN)による画像認識等の成果はめざましいが，人間の目には検知できないような摂動ノイズを加えることで，見た目とは全く異なる対象物に誤認識させる敵対的サンプル攻撃（AE: Adversarial Example attack） が深刻な問題として取り扱われている．本論文は，AE 攻撃を高精度で検出できる手法の提案を通して，実社会に貢献することを目指したものである．

本論文は，“Automated Detection System for Adversarial Examples on Images with Image Transformation and Filters”と題し，4 章と付録からなる．

第 1 章では，本研究の背景として課題の重要性と画像分析の基本手法について述べている．

第 2 章では，AE 攻撃について，幅広く既存研究を整理・分類している．

第 3 章では，本研究のコアとして，画像の幾何変換とフィルタリングによって AE 攻撃を高精度で検出する手法を論じている．幾何変換として，画像の回転(rot)と並進(trans) を適用し，フィルタリングとしてガウス平均法とメディアン(中央値)法を適用している．正常な画像であれば，これらの変換を加える前と後で，高確率で同じラベルに認識するが，AE 画像では，変換前と変換後で全く異なるラベルに認識されることを利用して，AE を検出する．従来法では，検知率が低かったり，変換後に正答ラベルに認識されないことが多かったが，Dang 氏の方法では，高確率で正しい画像ラベルを復元することができる．AE 攻撃手法は多数が提案されているが，Dang 氏は，特に強力なホワイトボックスアタックの中での代表的な 4 手法（FGSM，PGD，C&W および EAD 法）を，文字認識のデータセット MNIST と 100 万以上の画像を含む ImageNet に適用した．認識器には，Google Inception を含む数種類を用い，いずれの場合も高い精度で AE 攻撃を検知し，正答を復元できることを示した．従来法が，攻撃検知のための閾値を個別に設定しなければならないのに対し，本手法は，全自動で動作するという特長も有している．画像変換の角度や移動量，フィルタのサイズがパラメータとして残るが，一定の値ですべての画像に適用できる．最後に，これらの実験とシステム開発を通して，AE に許される摂動ノイズの大きさや量に関しても考察を行っている．

第 4 章では，結論として，本論文の提案についてまとめ，将来課題として，機械学習の強靭化研究の重要性を論じている．

Ⅱ. 論文審査結果の要旨

本論文では，AE について先行研究をよく調査し，AE の特性について徹底的な解析を行ったことがわかる．本論文で提案する 2 種類の AE 対策技術を組み合わせることによって，99.9%以上の確率で AE 画像を検出し，オリジナル画像の認識率も原画像の認識率に近い成績を保つことができている．これは，従来法に比べて非常に良い性能であり，多数の画像データベースと複数の攻撃手法に対抗できる，信頼性の高い技術を考案･実証できたと評価できる．

Dang 氏は，まず画像認識の定番とも言える DLNN に対する AE 攻撃を試み，実証することで，国際会議でのポスター発表を行い，ベストポスター賞を授与されている．幾何的画像変換による方法と，画像フィルタによる対策は，それぞれ査読付き国際会議(International Symposium on Computing and Networking，採択率 14.9%，Symposium on Cyberspace Safety and Security，採択率 26%）に採択されている．複数の DLNN に対する攻撃･対策法の研究からわかった AE 画像に加えるノイズの性質についての論文は，オンライン･ジャーナル（Intl. Journal of Advanced Computer Science and Applications，インパクトファクタ 1.324） に採択されている．さらに，幾何的画像変換による対策と画像フィルタによる対策を合わせた AE 検出方式をオンラインジャーナル（IEEE Access　インパクトファクタ 4.098）に投稿し，採択されており，情報学への貢献は大きい．

よって，本論文は，博士（情報学）の論文として合格と認められる．

Ⅲ. 審査経過

本審査委員会は，2020 年 1 月 22 日に口述試問を行い、その後、2 月 19 日に公聴会と最終試験審査を行った．審査に当たっては，博士学位のディプロマ・ポリシーに基づいて総合的に評価し，申請者が学位取得にふさわしい知見を持つものと判断した．