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

量子化誤差を考慮した Adversarial training の提案と評価
Proposal and Evaluation of Quantization Aware Adversarial Training

増田春樹 * 吉田康太 * 藤野毅 †
Haruki Masuda Kota Yoshida Takeshi Fujino

キーワード Adversarial examples, Adversarial training, Quantization aware training

あらまし

近年，深層ニューラルネットワーク(DNNs)を用いた画像認識技術の精度向上により，DNNs は自動運転車の周囲認識システムや防犯カメラの人物検知システムなど多くの分野に応用されている．

一方で，DNNs の入力に対し微小な摂動を加算することで意図的に誤認識を誘発させる Adversarial examples (AEs)攻撃が問題となっている．図１に手書き数字画像のデータセットである MNIST を用いた AEs の例を示す．MNIST を用いた通常時・攻撃時の分類精度の比較を図２に示す．図２の青色の線のように，未対策のモデルでは攻撃時に分類精度が著しく低下する．

DNNs の AEs に対する頑健性を高める訓練手法として，学習時に AEs を用いる Adversarial training (AT)[1]がある．図２の緑色の線のように，AT を用いることで攻撃時に分類精度の低下を抑えられることが分かる．しかし，エッジデバイスで DNNs を実行する場合はパラメータを 8bit に量子化する必要がある．ここで，32bit の AT モデルを後処理で 8bit に量子化する手法では，図２のグレー線のように AT の効果が低下してしまう．

そこで本稿では，量子化誤差を考慮しつつ DNNs の学習を行う Quantization aware training[2]と AT を組み合わせる手法を提案する．提案手法のイメージを図３に示し，分類精度を図２の黄色の線で示す．結果から，本手法によって AT モデルを後処理で量子化する場合(グレー)よりも分類精度が向上し，32bit で訓練した AT モデルと同等の頑健性を持つ 8bit パラメータの DNNs を訓練できた．

図１：MNIST を用いた AEs の例（FGSM で作成）

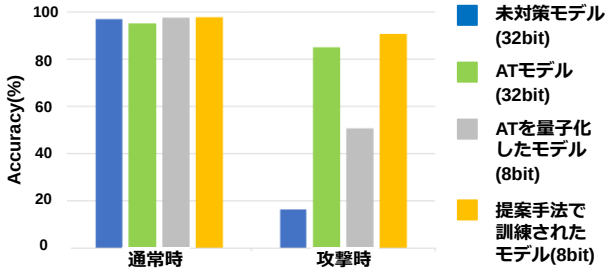

図２：MNIST を用いた各手法の分類精度の比較

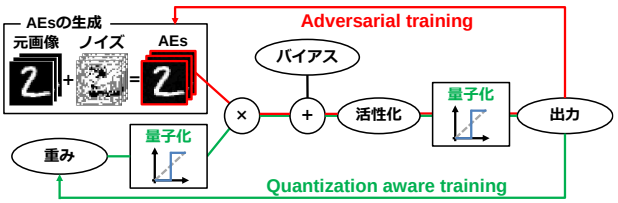

図３：提案手法のイメージ

参考文献

[1] Aleksander Madry et. al, "Towards deep learning models resistant to adversarial attacks", The International Conference on Learning Representations (ICLR), 2018

[2] Benoit Jacob et. al, "Quantization and training of neural networks for efficient integer-arithmetic-only inference", the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018

* 立命館大学大学院理工学研究科，〒 525-8577, 滋賀県 草津市 野路東 1-1-1,
{ri0068rr, ri0044ep}@ed.ritsumei.ac.jp.

† 立命館大学理工学部, 〒 525-8577, 滋賀県 草津市 野路東 1-1-1,
fujino@se.ritsumei.ac.jp.