未来社会創造事業　探索加速型
「超スマート社会の実現」領域
年次報告書(探索研究期間)

令和 2 年度
研究開発年次報告書

令和元年度採択研究開発代表者

［研究開発代表者名：藤野　毅］

［立命館大学　理工学部・教授］

［研究開発課題名：エッジ AI のハードウェアセキュリティに関する研究］

実施期間 ： 令和 2 年 4 月 1 日～令和 3 年 3 月 31 日

§1. 研究開発実施体制

(1)「立命館大学」グループ

①研究開発代表者:藤野　毅 (立命館大学理工学部, 教授)

②研究項目

(A) 攻撃対象となるエッジ AI エンジンの FPGA 実装/MPU 実装

(1) シストリックアレイを用いた AI エンジンの FPGA 実装

(3)サイドチャネル攻撃対策技術の実装

(B)エッジ AI ハードウェアに対する網羅的な攻撃手法と対策手法の研究

(1)サイドチャネル情報を用いた AI モデルのリバースエンジニアリング攻撃と対策

(3)サイドチャネル情報を用いたモデル抽出攻撃と対策

(4)イメージセンサインターフェースへの Adversarial Examples 攻撃と対策

(5)エッジ AI デバイスへのフォルト攻撃と対策

(C)AI 自身のセキュリティ全般に関する研究キャッチアップ

(2)「三菱電機」グループ

①主たる共同研究者:中井　綱人 (三菱電機　情報技術研究所, 研究員)

②研究項目

(A) 攻撃対象となるエッジ AI エンジンの FPGA 実装/MPU 実装

(2)AI エンジンの MPU を用いた実装

(3)サイドチャネル攻撃対策技術の実装

(B) エッジ AI ハードウェアに対する網羅的な攻撃手法と対策手法の研究

(2)サイドチャネル情報を用いた Adversarial Examples 攻撃と対策

§2. 研究開発実施の概要

2019 年度に FPGA および MPU 上に実装した深層学習(DNN)処理エンジン(研究項目 A)に対して, 演算時の消費電力を用いて, 学習パラメータを窃取する Model Reverse Engineering 攻撃 (研究項目 B-1), および, 演算時の処理時間を用いて Adversarial Examples を生成する攻撃手法の研究を行った(研究項目 B-2). これらの成果を国際会議での発表ならびに学術論文誌[*1,2]にまとめた.

新しい DNN 処理エンジンに対する攻撃手法として, 2種類のフォルト攻撃の研究を行った. 1つめは, イメージセンサとマイコンをつなぐ MIPI インターフェースに, 外部より不正なフォルト信号を注入することで, 画像内の特定の位置にノイズを安定して生成し, バックドア攻撃のトリガとして機能させる研究を行った(研究項目 B-4). また, MPU 実装した DNN 処理エンジンに対して, 異常なクロックすることで特定の命令をスキップさせ, DNN 処理結果を意図的に誤らせるという攻撃の研究を行った(研究項目 B-5).

研究項目 C に関しては，3つの成果を得た．第 1 に Adversarial Examples 攻撃対策として，ランダムノイズを画像に印加しさらに深層学習処理前に Denoising Auto Encoder 処理を行うという新しい手法を開発した．第 2 に DNN 演算におけるモデルパラメータの秘匿並びに改ざん防止を実現するために，セキュアな実行環境として TEE（Trusted Execution Environment）を使用することが今後主流となると予測され，従来よりもメモリ使用量が少なく実行時間の短い方式を提案し，ARM/Trust Zone 上に実装を行った．第 3 に，TEE 環境で行う DNN 演算の量を削減するため，DNN 処理前半の特徴抽出層は通常環境でアクセラレータを用いる方式の提案が行われているが，特徴抽出層のパラメータが攻撃者に知られると Adversarial Examples の生成が可能になることを確認した．これら3種の成果は本格研究におけるセキュリティを強化した DNN 処理ハードウェアに反映する予定である．

[*1]Tsunato Nakai, Daisuke Suzuki, Fumio Omatsu, Takeshi Fujino,” Adversarial Black-Box Attacks with Timing Side-Channel Leakage”, IEICE Tras. Vol.E104-A, No.1,pp.143-151.(2021)

[*2]Kota Yoshida, Mitsuru Shiozaki, Shunsuke Okura, Takaya Kubota, Takeshi Fujino, ”Model Reverse-Engineering Attack against Systolic-Array-Based DNN Accelerator Using Correlation Power Analysis”, IEICE Tras. Vol.E104-A, No.1,pp.152-161.(2021)