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
終了報告書(探索研究期間)

令和3年度
研究開発終了報告書

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　：　令和元年 11 月 1 日～令和 4 年 3 月 31 日

§1. 研究実施体制

(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. 研究実施の概要

【目的】Society5.0 で描かれるサイバー空間とフィジカル空間が融合された世界では, フィジカル空間の IoT 機器で得られたデータがサイバー空間のクラウドで AI を活用して解析され, その結果がフィジカル空間にフィードバックされる. ただし, 画像情報などの大容量のデータをすべて IoT 機器からクラウドに転送することは, 以下の 3 点から問題であり, IoT 機器内部で AI 処理を行う「エッジ AI」が必須である.

(1)通信量の増大:通信にかかる電力コストが大きく, 災害(通信途絶)時は動作不能となる

(2)低遅延動作:自動運転制御用車載カメラ搭載 AI など即応する必要のある用途では使用できない

(3)個人情報のプライバシー保護:監視カメラなどで生画像の漏洩リスクがある

特に重要な推論を行うためには, セキュリティを十分に考慮したエッジ AI の実装を行わなければならない. 例えば, AI 演算のパラメータ等は大量のデータを収集し, 多大な計算資源を使って作成した知的財産であるため, 保護する必要がある. また, エッジ AI のセキュリティ対策を考える場合, 従来の IoT 機器に対するセキュリティ対策に加えて AI 処理特有の攻撃を考慮する必要がある. AI 処理特有の攻撃とは, 大別すると以下の通り 4 種ある.

(1) 回避攻撃(Adversarial Examples 攻撃):AI 推論時の入力に微小なノイズなどの摂動を印加した

Adversarial Examples を AI に入力することで意図的に推論結果を誤らせる. 特に AI モデルが攻撃者に知られている時に脅威が増大する.

(2) ポイズニング攻撃:訓練データのごく一部を改ざんしモデルの訓練を行うことにより, モデルの推論精度を低下させたり, 攻撃者の意図する誤認識を誘発する. エッジ AI デバイスのモデルが不正に更新できる場合に脅威が増大する.

(3) モデル反転攻撃:学習済の AI モデルを解析することで, AI モデルが学習に使用したデータを解析・復元する. 訓練データのプライバシー保護に対する脅威となりうる.

(4) 抽出攻撃:正規の AI モデルでの多数回の推論結果を使用することで, AI モデルそのもの, もしくはこれと同等の性能を達成する別の AI モデルを入手する. AI モデルの知的財産に対する脅威となる.

本研究の目的は, 上記(1)～(4)の AI セキュリティの研究を含む研究を行うとともに, さらに攻撃者がエッジ AI 機器に物理的に接触できることを考慮した新しい攻撃手法と対策の研究を行うことである.

【成果】

代表的な成果を以下 7 件記載する.

(1) エッジ AI が動作しているときの演算処理時間というサイドチャネル情報を用いることで, 攻撃者が AI モデルを入手できない場合にも, 効率的に Adversarial Examples を生成できる新しい攻撃手法を提案した[*1].

(2) ポイズニングされた AI モデルに, 蒸留技術を使用することで誤動作誘発を防止し, さらにポイズニングデータを特定する防御手法を提案した.

(3) エッジ AI が動作しているときの消費電力波形を用いて, 内部で使用している AI パラメータを取得することが可能であるという原理的な実験を行った. モデル抽出攻撃と異なり正規のデバイスでの多数回の推論を行わずに, 正規のモデルを取得できるという知財に対する脅威となりうる[*2].

(4) エッジ AI が動作しているときの消費電力波形をモニタしながら, 適切なタイミングで, LSI に不正なクロック波形を入力することで, 特定のクラスの推論を行うことができることを実証した. [*3]

(5) エッジ AI で画像分類を行っている際に, カメラとエッジ AI のインターフェースに不正な電気信号を印加することで, 安定してポイズニング攻撃を発動させるという新しい攻撃手法を実証した. [*4]

(6) AI 演算をセキュアに実行する環境として TEE を検討し, AI 演算すべてを TEE 環境で実行しないと Adversarial Examples を生成できること, TEE で使用できる小規模なメモリサイズでも性能を低下させない演算手法を提案した.

(7) 本研究の成果を含む解説論文を「エッジ AI のハードウェアセキュリティ」というタイトルで電子情報通信学会の Fundamental Review 誌(2021 年 10 月号)にまとめた.

[*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)

[*3]Yuta Fukuda, Kota Yoshida, Takeshi Fujino, "Fault Injection Attacks Utilizing Waveform Pattern Matching against Neural Networks Processing on Microcontroller Date of Evaluation", IEICE Tras. Vol.E105-A, No.3, pp.300-310.(2022)

[*4]Tatsuya Oyama, Shunsuke Okura, Kota Yoshida, Takeshi Fujino, "Fault Injection Attacks Utilizing Waveform Pattern Matching against Neural Networks Processing on Microcontroller Date of Evaluation", IEICE Tras. Vol.E105-A, No.3, pp.336-343.(2022)