サイバーとフィジカルの高度な融合に向けたAI技術の革新

エッジAIのハードウェアセキュリティに関する研究

研究開発代表者： 藤野　毅　立命館大学　理工学部　教授

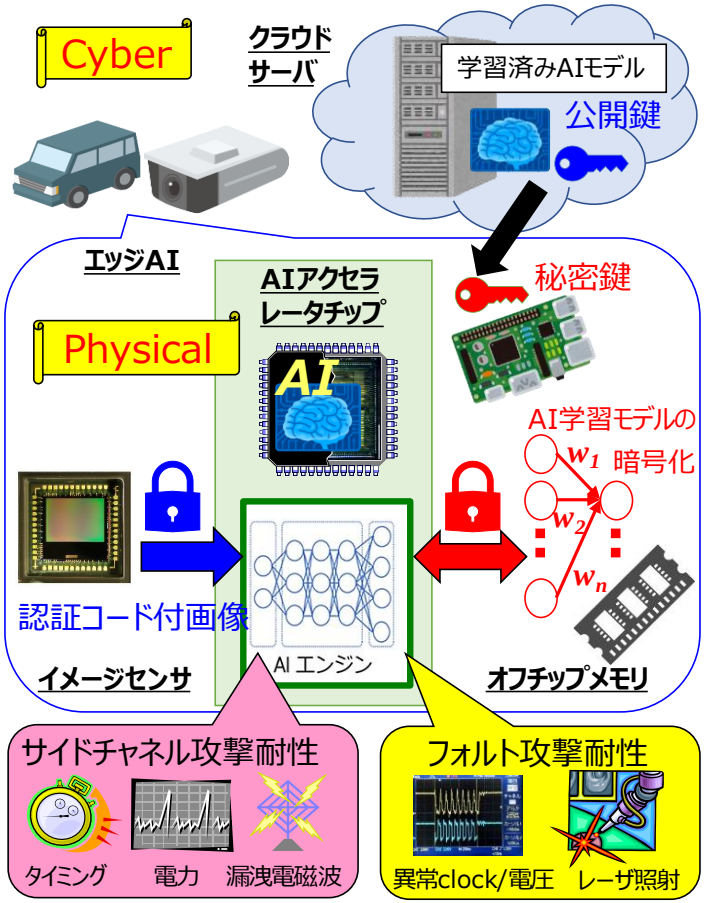
共同研究機関： 三菱電機　情報技術総合研究所

目的：

・エッジAIに搭載されるAIモデル（ネットワークの構造とパラメータ）の知的財産を守る！
・AI特有の誤動作（Adversarial Examples 等）を誘発する攻撃を防ぐ！
・AI学習時の画像データやAI推論時の画像データに含まれるプライバシー情報を守る！

研究概要：

スマートスピーカーに代表されるように，AI推論処理は主としてクラウドなどのサイバー空間で行われている．しかしながら，リアルタイム動作やプライバシー保護の観点から，フィジカル空間に配置されたAIで推論を行う「エッジAI」が必要となってきた．車載カメラや監視カメラ用途などの人命やプライバシー情報を扱う用途では，悪意ある攻撃による誤動作や機密情報の漏洩が許されない．一方，「エッジAI」は，攻撃者がAI処理ハードウェアに物理的に接触可能であることから，サイバー空間上に配置したAIよりも脆弱である．例えば，メモリバスデータの読出しや動作時の消費電力や漏洩電磁波の計測でAIモデルを窃取したり，異常クロック等の注入により誤動作の誘発を行うことが可能である．これらの攻撃を防止するためエッジAIにおけるハードウェアセキュリティの脅威と対策を研究する．AIモデルを秘匿し誤動作を防止する手段として，AIアクセラレータのメモリバスの暗号化，AIエンジンのサイドチャネル攻撃対策，フォルト攻撃対策，センサ情報の改ざん防止対策などのハードウェアのセキュリティ対策を行う．これらの対策が施されたセキュアなエッジAIを車載カメラや監視カメラに適用することで，AI社会の安心・安全に寄与する．



Innovative AI technologies for sophisticated integration of cyber and physical world

Research of Hardware Security on Edge AI Hardware

Project Leader : Takeshi Fujino
Professor, Department of Electronic and Computer Engineering, Ritsumeikan University

R&D Team : Information Technology R&D Center, Mitsubishi Electric Corporation

Summary :

"Edge AI" is getting more important from the viewpoint of real-time operation and privacy protection. On the other hand, "edge AI" on the physical space has more security vulnerability than "cloud AI" on the cyberspace, because attackers can physically access AI processing hardware. For example, AI models (DNN network structure and weight parameters) can be revealed by memory-bus tapping or side-channel attack exploiting power consumption or electromagnetic waves. Malfunction can also be induced by fault-injection attack. In this research, hardware security threats on "edge AI" are comprehensively analyzed, and countermeasures are studied to prevent these attacks.

<Security Countermeasures> (1) Data encryption of AI models against memory bus tapping. (2) Elimination of Side-channel leaks and increase of fault tolerance on AI processing hardware. (3) Integrity check and/or encryption from CMOS image sensor data.

<Benefits> (1) Protection of intellectual property on AI models. (2) Prevention of AI-specific malfunctions (such as Adversarial Examples). (3) Privacy protection of image data used for AI learning or inference.