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

イメージセンサインターフェースへのフォルト攻撃でトリガするDNNへのバックドア攻撃

Backdoor Attack against DNN triggered by Electrical Fault Injection on Image Sensor Interface

大山　達哉 * 大倉　俊介 * 吉田　康太 * 藤野　毅 *
Tatsuya Oyama Shunsuke Okura Kota Yoshida Takeshi Fujino

キーワード イメージセンサインターフェース，MIPI，フォルト攻撃，バックドア攻撃， DNN

あらまし

DNN(Deep Neural Network) による画像認識システムへの攻撃として訓練データのポイズニングを用いたバックドア攻撃が知られている [1]．特定のマーク (トリガマーカ) が画像内の所定位置に配置されている際は，誤った (ターゲット) ラベルに分類するように DNN を学習させる．攻撃者によって撮影画像の所定位置にトリガマーカが配置されるとバックドアが起動し誤分類を誘発するが，トリガマーカの位置や大きさが撮影環境で変化するため，安定してバックドアを起動するのは難しい．

図 1 のように，著者らは SCIS2021 でイメージセンサとホストデバイスの間で画像情報の通信する MIPI(Mobile Industry Processor Interface) 上で電気的な信号を注入して画像上にトリガマーカを生成するフォルト攻撃を提案し，安定してバックドアをトリガできることを示した．昨年の報告ではグレイスケールの低解像度手書き文字認識を行う DNN に攻撃を行ったが，本稿ではより現実的なシナリオとしてカラーの道路標識画像分類 DNN を対象に同様の攻撃とトリガマーカの解析結果を報告する．

実験では，図 2(a) の実験環境の下，Raspberry Pi とカメラモジュール間の MIPI 信号に 2 系統の攻撃用の小振幅差動信号を注入した．2 系統の信号が逆相の場合には正常な信号が送信され，同相の場合には攻撃信号が送信されることを利用して，画像内の特定領域にトリガマーカを付加した．図 2(b) に示した実験結果例では，MIPI にフォルト攻撃せずに撮影を行った場合に STOP に分類される標識が，トリガマーカによって速度制限 120km/h に分類されたことが確認できる．また，トリガマーカはドット模様で一部背景が反映された色となっているが，これらの現象に関して考察した結果を述べる．

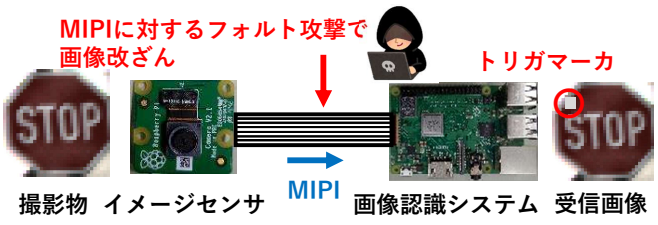

図 1: トリガマーカを撮影画像に付加する提案手法

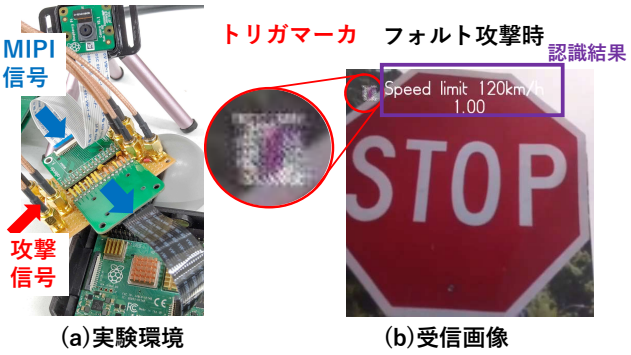

図 2: 実験環境 (a) とフォルト攻撃時の実験結果例 (b)

参考文献

[1] Tianyu Guet, al,“’ BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain’ vol.abs/1708.06733, 2017.

* 立命館大学, 〒 525-8577 滋賀県草津市野路東 1-1-1, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga, Japan