5ZA-05

IoTマルウェア画像分類手法に対する実行可能なノイズ付与による攻撃手法の検討

川田 隼大† 稲村 浩† 石田 繁巳†
†公立はこだて未来大学

1 はじめに

現在，IoT デバイスの脆弱性を利用したマルウェアが増加している．Zscalar が 2023 年度に発表した IoT の脅威レポートによると，2022 年度に比べ IoT マルウェア攻撃の数は 400%以上増加している [1]．このため，大量の IoT マルウェアの動向を正確かつ速やかに把握し，対策を打つことが重要である．

大量の IoT マルウェアの動向を把握するには，マルウェアの分類が役立つ．Su らはマルウェアのバイナリを画像化し，深層学習を用いて分類することによって従来よりも高い精度かつ高速な分類を実現した [2]．しかし，画像化を用いた手法は画像へのノイズ付与によって，正確な分類を妨げることができる．

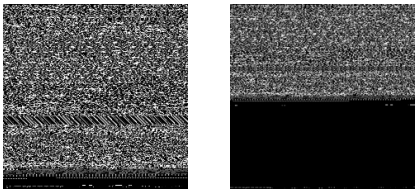
画像へのノイズ付与を行った研究として，自然画像分類を対象としたものが数多くある．Su らは画像の任意の 1 ピクセルを変更し，約 20%の分類精度低下が可能であることを明らかにした [3]．Android のマルウェア分類器を対象としたものもある．Gu らはヘッダーテーブル部分を除外した範囲でノイズを付与し，最大 81.34%の攻撃成功率を達成した [4]．

関連研究 [3, 4] をはじめ従来の研究では，画像へのノイズ付与で分類精度の低下を実現しているが，実行可能性の維持については議論が少ない．実行可能性が維持されていない検体は，マルウェアとして機器に被害を及ぼすことはなく，解析者にとって分類の必要性がない．このため，IoT マルウェアの画像化を用いた分類手法を対象とした場合は，実行可能性を維持したノイズ付与を行う必要がある．

そこで本研究では，生成した IoT マルウェアが実行可能性を維持できるように画像へノイズを付与する攻撃手法の検討を行う．

2 空セクションの追加

本研究では，実行可能性を維持した画像へのノイズ付与による攻撃手法として，空セクションの追加を提案する．空セクション追加前後の IoT マルウェア画像



(a) 追加前 (b) 追加後

図 1: 空セクション追加前後の IoT マルウェア画像

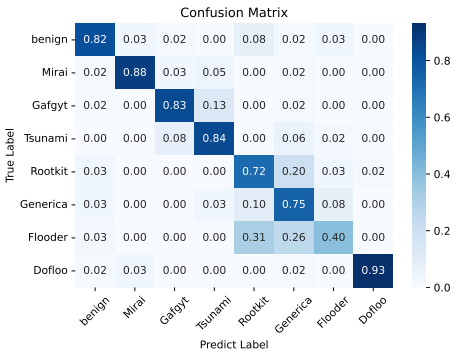


図 2: ベースモデルの混同行列

を**図 1**に示す．空セクションのデータは全てゼロとなっており，輝度 0 のピクセルに変換されるため，グレースケール画像における黒の領域が増加している．

空セクションは，実行時に動作しない冗長なセクションであり，通常は ELF ファイルのリンク時に削除や最適化が行われる．このため空セクションの追加は，バイナリ変更による画像へのノイズ付与を実現しつつ，マルウェアの実行可能性を維持していると言える．

3 評価

3.1 ベースモデルの作成

本研究では IoT マルウェアを画像化し，マルウェアファミリーに分類するベースモデルの作成を行った．

データセットは Olsen らによるオープンソースデータセット [5] のうち，検体数の差が最も少ない Intel80386 アーキテクチャのものを使用した．本研究では，検体数の差による分類の精度低下を避けるため，全てのファミリから 200 検体をランダムに抽出したものと，200

An Attack using Executable Noise in IoT Malware Classification
Hayata Kawata†, Hiroshi Inamura†, Shigemi Ishida†
†Future University Hakodate, Japan
†{b1021243, inamura, ish}@fun.ac.jp

表 1: 分類精度の変化

ファイルサイズ比率	分類精度
0% (Original)	79.05%
20%	64.32%
40%	56.55%
60%	47.33%
80%	41.75%

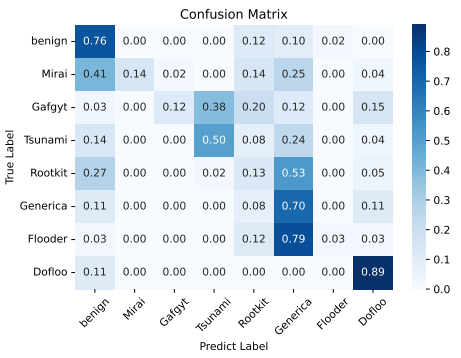


図 3: 空セクション追加後の混同行列

検体に満たないマルウェアファミリーである Flooder の 118 検体を IoT マルウェア画像分類器の作成に用いるデータとした．ベースモデルは VGG16 をファインチューニングすることで作成し，学習時のパラメータは，エポック数を 40，最適化関数を Adam，バッチサイズを 6，学習率を 0.00001 とした．

作成したベースモデルの評価を行うため，テストデータの分類を行ったところ，分類精度は 79%であった．テストデータの分類結果の混同行列を**図 2**に示す．Flooder を除く 7 つのクラスにおいて 70%以上の分類精度であった．200 という少ない検体数でも，各マルウェアファミリの画像特徴を捉えられていることがわかる．

3.2 評価結果

テストデータに対し，ファイルサイズの 20%，40%，60%，80%の空セクション追加を行った．空セクションを追加した IoT マルウェア画像の分類精度を**表 1**に，ファイルサイズの 80%の空セクション追加を行ったデータの分類結果の混同行列を**図 3**にそれぞれ示す．追加量の増加に伴って分類精度が低下しており，最大 38%の低下が見られた．これにより，空セクションの追加のみでも画像化を用いた分類手法への単純な攻撃として一定の効果を持ち得ることが確認できた．

表 2: 良性への誤分類率の変化

ファイルサイズ比率	良性への誤分類率
0% (Original)	0.04%
20%	0.09%
40%	0.10%
60%	0.12%
80%	0.15%

空セクションを追加した IoT マルウェア画像の良性への誤分類率を**表 2**に示す．追加量の増加に伴って，マルウェア検体における良性への誤分類率増加が見られた．良性への誤分類率が増加した原因として，空セクション追加後の画像のリサイズが影響していると考えられる．モデルの学習に使用した良性のデータセットは，共通した画像特徴を持っていない．このため，ベースモデルが画像特徴を読み取れなかった場合，良性へ誤分類する．空セクション追加後の画像のリサイズによって，判断根拠となる画像特徴の細部が失われたため，ベースモデルは画像特徴を捉えられなかった可能性がある．

4 おわりに

本稿では，IoT マルウェアの画像化を用いた分類手法への実行可能性を維持した攻撃手法として，画像への空セクションの追加を提案し，有効性を評価した．評価の結果，空セクションの追加によって，分類精度が最大 38%低下し，リサイズによる良性への誤分類増加が明らかになった．これにより，空セクションの追加が画像化を用いた手法に対して一定の攻撃効果を持ちうることを示した．

参考文献

[1] ThreatLabz: 2023 年版 Zscaler ThreatLabz エンタープライズ IoT および OT の脅威レポート— Zscaler, `https://bit.ly/3UV576b` (2023). (Accessed on 12/16/2024).

[2] Su, J., Vargas, D. V., Prasad, S., Sgandurra, D., Feng, Y. and Sakurai, K.: Lightweight Classification of IoT Malware Based on Image Recognition (2018).

[3] Su, J., Vargas, D. V. and Sakurai, K.: One Pixel Attack for Fooling Deep Neural Networks, *IEEE Transactions on Evolutionary Computation*, Vol. 23, No. 5, pp. 828–841 (2019).

[4] Gu, S., Cheng, S. and Zhang, W.: From Image to Code: Executable Adversarial Examples of Android Applications, *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*, Tianjin China, pp. 261–268 (2020).

[5] Olsen, S. H. and OConnor, T.: Toward a Labeled Dataset of IoT Malware Features, *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, Torino, Italy, pp. 924–933 (2023).