Capsule Neural Networkに基づくマルウェア分類手法

Malware classification method based on Capsule Neural Network

加藤 秀記・マネジメント分科会・情報セキュリティ大学院大学

In recent years, the number of malware has been increasing rapidly. Then, the existing method for analyzing the malware is performed manually, and there is a problem that it takes time to analyze.
It is extremely important to establish an efficient analysis method in order to take appropriate measures against malware threats that are expected to increase in the future, and malware classification plays an important role as one of the methods.
In this study, malware is classified using a new deep learning method for image classification called Capsule Neural Network, and its superiority is clarified by comparing it with the classification method using convolutional neural networks that has been widely used so far. Furthermore, we proposed guidelines that lead to improved accuracy and learning time when creating learning models using Capsule Neural Network.

背景

マルウェアの現状

2019年時点で、マルウェアの数は10億を突破！
近年はIoT機器をターゲットにしたマルウェアも出現

今後も増加傾向

https://www.av-test.org/en/statistics/malware/

既存マルウェア対策の課題

これまでのマルウェアの解析方法

- 静的解析
- 動的解析

既存マルウェア解析の問題点

- 高い技術力、経験などが必要
- 人手のため、解析に時間がかかる

マルウェア解析の効率化が急務！！

マルウェアの分類について

マルウェア解析の中で、**どの種類に属するか**の情報は重要！

WHY?

同種類の既存マルウェアとの比較が可能！
→ノウハウの活用から、効率的な解析が可能になる

高精度・短時間でのマルウェア分類手法の確立はマルウェア解析の効率化に重要な役割を持つ！

課題・研究目的

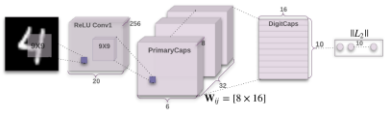
機械学習を用いたマルウェア分類研究

マルウェアを画像化し、模様のパターンから分類[1]
→画像分類で高い精度を持つニューラルネットワーク(CNN)を使う研究がある。

近年CNNにも欠点があると指摘

新たな学習モデルの誕生

Capsule Neural Network(CapsNet)[2]



最も大きな特徴：Capsuleによるネットワーク
Capsule：neuronをまとめてベクトル化した単位

CapsNetの強み

高精度・高汎用性を実現するにあたっての、
学習用のデータセットの指数関数的な増強が不要。
→限られたデータでも高い精度を維持できる可能性あり

近年、CapsNetも徐々に研究事例が報告され、マルウェア分類に利用した研究もある

過去研究における課題と研究目的

課題1

CapsNetによる分類の精度は95%以上で高精度
→ただし、CNNでも同程度の精度

精度比較だけではCapsNetによるメリットは少

課題2

学習モデルの作成で、画像サイズ、Capsuleの次元数などの条件の選定理由がない
→**最適条件はTry&Errorの繰り返しで見つけるしかない**

モデル作成が非効率

研究目的

目的1
精度以外のメリットの発見(過学習耐性)

目的2
モデル作成におけるガイドライン作成

結果・考察

過学習への耐性について

モデル	Accuracy	過学習耐性
CNN	高い	低い
CapsNet	高い	高い

CapsNetがCNNより高い過学習耐性を示した

情報が**ニューロンのスカラー値からCapsuleのベクトル**になることで、保持できる情報量が増加したことが考えられる。

モデル作成のガイドライン

データセット作成

画像から抽出できる特徴の数が多いほど、Accuracyの向上につながる。そのため、学習に使用する画像のサイズは大きいほどよい。

第1層目の畳み込みとReLUのチャンネル数

学習に使用できる特徴の量がAccuracyにおいて重要な要素となる。よって、第1層目の畳み込みとReLUのチャンネル数は多いほどAccuracyの向上につながる。

第2層目のPrimaryCapsの奥行(次元数)

PrimaryCapsの奥行(次元数)は、増加させることである程度の学習時間の短縮が可能である。精度に影響なく学習時間を短縮したい場合は、次元数を増加させる。

今後の研究指針

- 本実験で得た傾向の根拠を明確化
- 過学習以外(F1値、Recallなど)の優位性の明確化
- 学習の入力として、画像以外の方法の模索(※)

(※)先行研究としてWord2vecとCNNを組み合わせたText Classificationを実現する学習モデルが考案されている。
→CapsNetとの組み合わせを検討。

参考URL

[1] : Malware images: visualization and automatic classication, VizSec, p. 4 (2011). Nataraj, L. et al.
[2] : Dynamic routing between capsules. Sara, S., Nicholas, F., Geoffrey, H. Advances in Neural Information Processing Systems, pp. 3859–3869 (2017)