SecHack365

マルウェア検出手法に対する汎用敵対的摂動

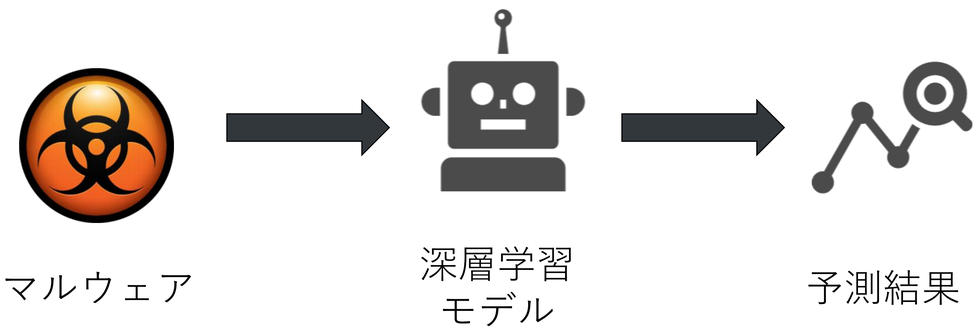
研究駆動コース　宇野 貴士

背景

Background > end2end.sec

End to End型マルウェア検出器

マルウェアの増加傾向が指摘されている
マルウェアのバイナリを直接入力として用いる
End to End型マルウェア検出器が注目を集めている

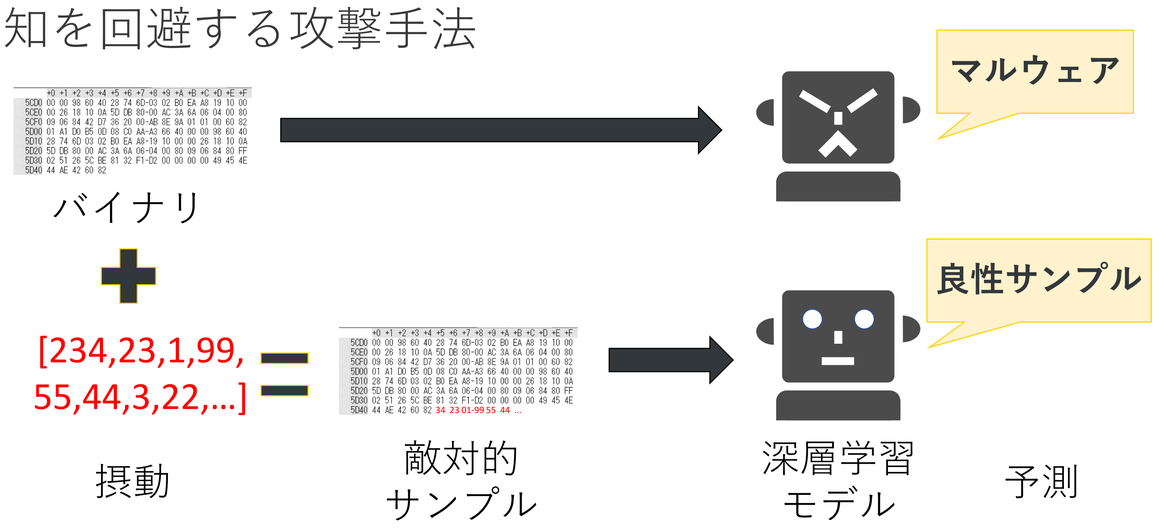


End to End型の学習器には敵対的サンプル用いた攻撃に対して脆弱なことが懸念される

Background > adversarial_attack.hack

マルウェア検出に対する敵対的攻撃

マルウェア検体を良性サンプルと誤認させるための摂動（僅かなノイズ）をバイナリに加えることで検知を回避する攻撃手法



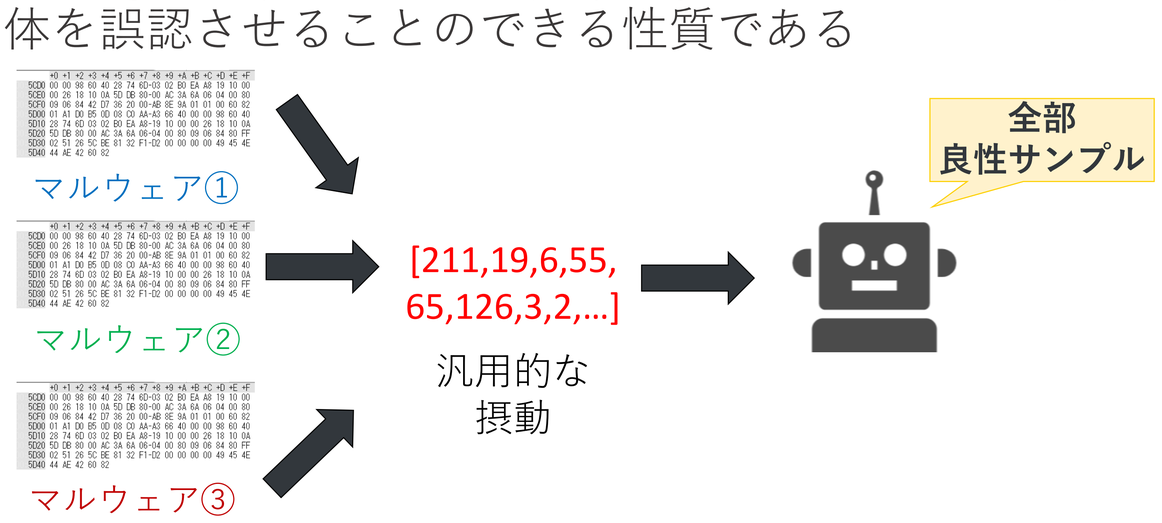
目的

Purpose > UniversalAdversarialPerturbation.365

汎用的な摂動の生成

End to Endマルウェア検出器を対象に、汎用的な性質を持つ摂動を計算することを目的とする

汎用的とは単一の摂動で非常の多くのマルウェア検体を誤認させることのできる性質である



[1] Raff, E., Barker, J., Sylvester, J. et. Al. : Malware detection by eating a whole exe. AAAI Workshops, 2018.

提案主砲

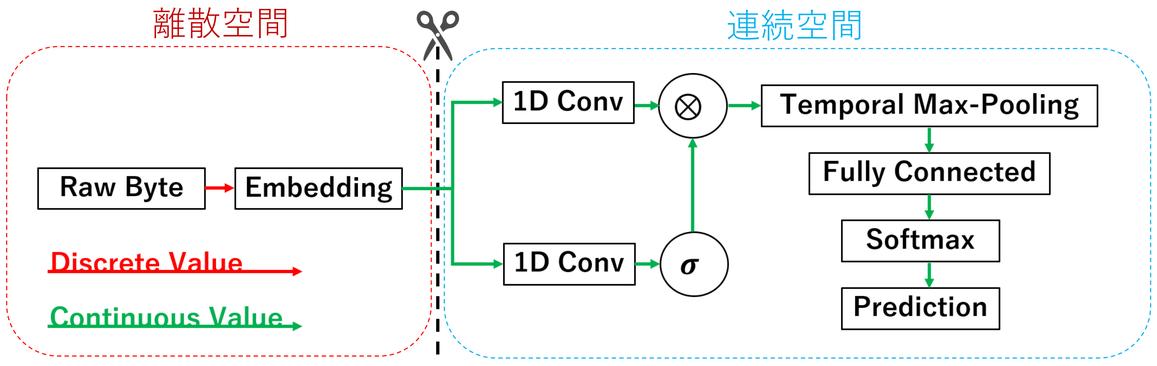
もしかして ： 提案手法

Propose > attack_method.ytel

マルウェア検出器に対する汎用的摂動生成

既存の攻撃手法は、連続値を対象としている
そこで攻撃対象モデル[1]を分割し中間表現を入力として扱い、連続空間において汎用的摂動を構築する

次に得られた、汎用的摂動の中間表現からバイナリを再構成することにより、汎用的摂動を取得する



Propose > attack_method.suberu

摂動の量産手法

単一の摂動ではパターンマッチングで防衛可能
そこで、中間表現にてランダムなノイズを与えてから再構成することで、元となる摂動と同様の性質を有しつつも、異なるバイナリを持つ摂動を量産する手法を提案する

実験

Experiment > generate_perts.thx

実験結果

生成した摂動のエラー率、信頼度減少率を測定した

高いエラー率により、非常に多くの検体を誤認させる性質を有していることが分かる

	提案手法				ランダムノイズ
検体数	5000	1000	500	250	
エラー率	55.1%	56.0%	53.4%	40.8%	8.2%
信頼度減少率	62.7%	63.0%	65.7%	55.4%	19.0%

量産した摂動の性能

ノイズ強度	0.000	0.010	0.050	0.100	0.500
エラー率	56.0%	52.6%	52.4%	42.4%	23.4%
バイナリ一致率	100%	87.4%	48.2%	20.4%	1%

まとめ

深層学習を用いたマルウェア検出器が汎用敵対的摂動に対して脆弱である可能性を示した

情報処理学会論文誌採録 、Vol.62 6月号にて掲載予定

対象としてた識別器以外でも提案手法が有効か検証する必要がある

国立研究開発法人 情報通信研究機構