情報処理学会論文誌 Vol.62 No.6 1368–1372 (June 2021)

テクニカルノート

深層学習により構築されたEnd to Endマルウェア検出器に対する汎用敵対的摂動生成手法

宇野 貴士[1] 宮尾 秀俊[1,a)] 丸山 稔[1,b)]

受付日 2020年12月15日, 採録日 2021年2月2日

概要：近年の深層学習の発展は著しく，マルウェア検出における静的解析手法としても応用が模索されており，深層学習によるマルウェア検出の精度向上，コスト削減が期待されている．一方で，画像・音声をはじめとした認識技術において指摘されているように，深層学習によるマルウェア検出も敵対的サンプル（Adversarial Examples）に対して脆弱であることが懸念されている．本論文では汎用的な敵対的摂動の生成手法に着目し，既存手法と比べより制限された実験環境における敵対的サンプルの構築手法を提案し，マルウェア検出における脆弱性について検証する．

キーワード：深層学習，セキュリティ，敵対的サンプル，マルウェア検出

Universal Adversarial Perturbation against End to End Malware Detector

Takashi Uno[1] Hidetoshi Miyao[1,a)] Minoru Maruyama[1,b)]

Received: December 15, 2020, Accepted: February 2, 2021

Abstract: It has been shown that deep learning can achieve very good performance various kinds of tasks including image classification, object recognition, machine translation etc. Recently, deep learning based malware detectors was proposed and showed promising performance. For applications of deep learning in security-related field, robustness against adversarial perturbation is required. To make a robust malware detector, analysis on existence and properties of adversarial examples is very important. In this article we investigate a method to make universal adversarial perturbations (UAP) for CNN-based malware detector. We show the UAP generated by our method could cause misclassification of the CNN-based malware detector.

Keywords: deep learning, security, adversarial examples, malware detector

1. はじめに

深層学習を用いて構築されたモデルはきわめて高い認識，生成能力があることが示され，様々な分野における State Of The Art（SOTA）を獲得しており，顔認証，自動運転，ロボット制御をはじめとするミッションクリティカル領域への応用が模索されている．この動向はセキュリティ分野へも波及しており，これまでは多大なコストとセキュリティ専門家の知見・探求によって実現されていた静的解析によるマルウェア検出を深層学習の持つ高い特徴抽出能力を生かし，効率的に行う試みがある．しかしながら，マルウェア検出問題を扱うデータセットの多くは，専門家が高度な知見と専門的経験則によって抽出した特徴量を配布しており，多大な人的資源が必要である．こうしたデータセットを対象とした，機械学習を用いたマルウェア検出手法は，既存の検出技術と同様に，多大な人的資源が必要であるという問題を有している．そうしたなかで，マルウェア検体の実行ファイルから直接的に特徴抽出を行う深層学習を用いた手法が提案され，高い精度で既存マルウェアの亜種であるゼロデイマルウェアを検出可能であることが示

[1] 信州大学大学院総合理工学研究科
Department of Electrical and Computer Engineering, Shinshu University, Nagano 380–0928, Japan
[a)] miyao@cs.shinshu-u.ac.jp
[b)] maruyama@cs.shinshu-u.ac.jp

されている [1], [2], [3].

一方文献 [4] によって，深層学習を応用した識別器と正しく識別可能な入力信号から，恣意的に誤認識を発生させる摂動[*1]が計算可能であることが示された．本来の入力信号と摂動を重ね合わせたものを敵対的サンプル（Adversarial Examples）と呼ぶ．敵対的サンプルは [5] によって深層学習を用いたマルウェア検出器においても計算可能であることが示されており，このような攻撃手法に関するいっそうの解析や防衛手法の構築が求められていくものと考えられる．特に Adversarial Training [6] に代表される識別器の敵対的サンプルに対する堅牢性向上による防衛手法の多くは，敵対的サンプル生成手法が既知である必要性がある．このため，敵対的サンプルの生成可能性を検証することは，深層学習によるマルウェア検出器の堅牢性を高め，社会実装していくうえで重要であるといえる．

文献 [5] ではマルウェア検出においても，モデル構造，パラメータが既知であるとき，各マルウェア検体ごとに敵対的サンプルを計算可能であることが示された．文献 [5] は重大なインシデントを引き起こすことが危惧されるが，攻撃者は学習済みモデルを入手する必要があるため，現実的に実施可能な攻撃シナリオを構築するのは難しい．一方で画像認識，音声認識，文章分類においては，特定の入力信号だけではなく，任意の入力信号に対して有効な敵対的摂動である Universal Adversarial Perturvation（UAP，汎用的敵対的摂動）[7], [8], [9] の存在が知られている．もし，マルウェア検出において，文献 [7] と同様の性質を有する汎用的な摂動を多種生成することが可能である場合，敵対的サンプルを用いた，実社会において実施可能な脅威度の高い攻撃シナリオが構築可能となる危険性がある．そこで，本論文においては文献 [7] をもとにし，様々なマルウェア検体にわたって，マルウェア検出器の誤動作を引き起こさせるような摂動を計算可能であるか検討し，その摂動生成アルゴリズムおよび性能に関して小規模なマルウェア検出器を用いて検証する．

2. 深層学習を用いたマルウェア検出器に対する敵対的サンプル生成関連手法

文献 [2] においてマルウェア検体のバイナリ値を直接入力とし，深層学習の持つ高い特徴抽出能力を生かし，専門的経験則を用いた方針決定を行わずに，End to End 学習可能な識別関数を構築する静的解析手法である MalConv [2] が提案された．MalConv [2] は高い精度を持ち，バイナリから直接的に特徴抽出を可能にする利点を持つが，同時にマルウェア検体を直接的に改編することによって構築される敵対的サンプルに対して脆弱である可能性が指摘されている [5]．本論文においては文献 [2] がどういった条件において堅牢な運用が可能であるかを明らかにするため，単純なホワイトボックス[*2]以外の問題設定における，敵対的攻撃可能な条件を調べる．

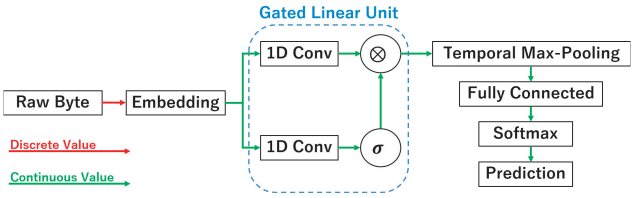


図 1 Malconv のモデル図
Fig. 1 Structure of MalConv.

2.1 MalConv

文献 [2] において提案された MalConv はバイナリファイルという長大な系列データの入力と大量のデータセットを用いた学習を両立するために，時系列データの処理に一般的に用いられる RNN ではなく，CNN をベースにしたモデルである．MalConv のモデル構造を**図 1** に示す．モデルの構造としては，バイナリファイルに含まれるバイト列を 0 から 255 までの離散空間における値として扱い，それを直接入力として用いる．モデル内では初めに埋め込み層により整数表現から実数表現へと変換する．その後，1 次元畳み込み層とシグモイド関数 σ から構成される Gated Linear Unit [10] により特徴抽出を行い Softmax 層による分類を行う．MalConv は十分な数のマルウェアサンプル（PE ファイル）によって学習した場合，94.0%の認識精度を獲得している [2]．この精度はヘッダーファイルのみを対象とする手法 [1] や，バイト列の完全一致による識別と比べ高い結果である．

2.2 MalConv に対する敵対的サンプル

MalConv は埋め込み層を除いて考えれば，一般的な画像認識に用いられる CNN と類似した構造を持つ．CNN は画像識別に関して優れた能力を有することが知られているが，このような識別器に誤認識を発生させる敵対的サンプルを生成できることも示されている [4], [6]．MalConv は画像識別と同様に CNN を使用するが，画像識別で検討されているような入力データ全体に摂動を添加する方式での敵対的サンプル生成は PE ファイルを実行不可能にする可能性が高いため不適切である．文献 [5] では MalConv に対する敵対的サンプル生成手法として，PE ファイルの性質に着目し，バイナリの終端部分にペイロードと呼ばれるバイト列を配置し，FGSM [6] を複数回実行することにより，強力な摂動を生成する Iterative-FGSM を用いて誤認識を発生させる方式を用いている．

[*1] 人間の感覚器官が知覚できないほど微細な信号

[*2] 学習済みの識別器とデータセットが開示されている条件設定における攻撃可能性テスト

2.3 Universal Adversarial Perturbation

Universal Adversarial Perturbation（UAP）とは汎用的（Universal）な性質を持つ摂動の生成手法である．敵対的サンプルにおける汎用的とは具体的には次の性質を有することである．

- 入力信号の出力結果のカテゴリに依存しない．
- 入力信号に対して依存しない．
- 特定の識別関数に対して依存しない（転移性）．

このとき，UAP が生成可能である場合，文献 [5] と比較し高い転移性を有するため，ブラックボックス[*3]な条件であっても敵対的サンプルが生成可能であることが示唆される．本論文においては，特定のサンプルと生成対象となるモデルの構造およびパラメータを参照する文献 [5] と比べ，特定のサンプルのみに有効ではないという点において量産性・転移性に優れた UAP の生成手法を考案し，その性質を検証する．

3. 提案手法

本論文においては End to End 学習可能なマルウェア検出器である MalConv [2] に対する UAP 生成手法を提案する．

3.1 離散空間と連続空間の分離

自然言語処理に代表される，離散値を入力として用いる深層学習モデルは一般に埋め込み層によって入力を実数ベクトルへと変換し，以降の処理は実数に対する処理を定義する．MalConv においてもこれは同様である．連続空間における敵対的サンプル生成は多くの関連研究が行われており，様々な問題設定において敵対的サンプルを生成可能であることが知られている一方，離散空間での敵対的サンプル生成は距離の定義のあいまいさなどにより研究は進んでいない．したがって本論文においては埋め込み層以前と以後においてモデルを分割し，連続空間における敵対的サンプルを生成し，それを埋め込み層の逆関数を定義し，離散値に戻す手法を用いる．

本論文においては MalConv を $\mathrm{F}(X_i)$ としたとき，その入力層および離散空間を $\mathrm{Em}(X_i)$，連続空間を $\mathrm{Pr}(x_i)$ とする．したがって $\mathrm{F}(X_i) = \mathrm{Pr}(\mathrm{Em}(X_i))$ という関係が成り立つ．また，埋め込み層の逆関数を $\text{re-Em}(x_i)$ とする．MalConv では各バイナリ値ごとに特徴ベクトルへと変換する．そのため，任意の特徴ベクトルに対して，最近傍法を用いることでバイナリ値を近似することが可能である [5]．したがって，関数 $\text{re-Em}(x_i)$ は容易に定義することができる．

[*3] 学習済みの識別器と学習用データセットに対し直接的に参照はできず，識別結果のみを取得可能な条件設定における攻撃可能性テスト

3.2 MalConv に対する汎用敵対的サンプル生成アルゴリズム

UAP 生成アルゴリズム [7] では十分な要素数を持つ入力信号群を用意し，それぞれの各入力信号（各要素）に対して有効に作用する摂動から，有効性を維持した平均的な摂動を探索する手法である．本論文においては [5] によって提唱されている，MalConv に対するペイロード型の Iterative-FGSM [6] 生成手法を応用し，各マルウェア検体に対する敵対的摂動を計算し，得られた敵対的摂動に平均化手法を適用する．

文献 [7] において提案される UAP 生成アルゴリズムをもとにし，ペイロード型へと変更を行った提案アルゴリズムを Algorithm 1 に示す．

ここで，出力値が 0〜1 の連続値であるため，識別カテゴリの境界値を β とする．Algorithm 1 のパラメータとして，目標エラー率を δ，各サンプルに対する摂動強度制御定数を c とする．また，マルウェア検体群を D とし，各マルウェア検体を X_i，マルウェア検体 X_i を特徴ベクトルへと変換した連続量を x_i とする．

4. 実験

Algorithm 1 によって提案した手法によって，MalConv における UAP 生成実験を行う．ただし，本論文においては文献 [2] の実験で使われた 2 MB までのシーケンス長の検体を入力可能な MalConv ではなく，約 80 KB までのシーケンス長に対応した小規模なモデルを用いる．このとき，VirusShare [11] にて配布されているマルウェア検体のうち約 30%は 80 KB 以下である．したがって，本論文における実験で用いる小規模モデルは文献 [2] の実験と比較した場合には小規模であるが，実用上において流通している多くのマルウェアを入力可能である．

Algorithm 1 Universal Adversarial Perturbation against Malware Detector

Input: **MalwareDataPoint** D, **Regressor Pr**,
Embending Em, **thres** β, **ControlPerts** c,
ObjectiveErrorRate δ
Output: **Universal Perturbation Vector** V
Init $v \leftarrow 0$
while $ErrorRate < \delta$ **do**
  $D' \leftarrow \mathrm{Shuffle}(Em(D))$
  for each datapoint $x_i \in D'$ **do**
    if $Pr(x_i + v) < \beta$ **then**
      $\triangle v_i \leftarrow v$+Iterative-FGSM($Pr$,$x_i + v$)$\times c$
      $v \leftarrow v + \triangle v_i$
    end if
  end for
  $ErrorRate = \mathrm{Err}(D, v)$
end while
$V = \text{re-Em}(v)$

表 1 提案手法およびランダムノイズのエラー率・信頼度減少率の比較
Table 1 Results table of error rates and confidence reduce rate for the proposed method and random noise.

		提案手法					ランダムノイズ
	検体数	5,000	1,000	500	250	100	
バイナリ	エラー率	0.551	0.560	0.534	0.408	0.221	0.082
	信頼度減少率	0.627	0.630	0.657	0.554	0.364	0.190
特徴量	エラー率	0.696	0.701	0.690	0.605	0.434	0.362
	信頼度減少率	0.713	0.723	0.733	0.674	0.558	0.471

4.1 実験環境

生成に用いる，実験環境およびパラメータを以下に示す．

マルウェア検出器 MalConv（学習済みモデル）

入力層 80,000 バイト

出力層 シグモイド関数

（0：マルウェア 1：グッドウェア）

マルウェア検体 VirusShare [11] より MalConv が十分に高い信頼でマルウェアと判定する PE ファイルを取得

生成用データセット 5,000 検体

検証用データセット 2,500 検体

パラメータ

しきい値 $\beta = 0.5$，摂動強度制御定数 $c = 0.2$，目標エラー率 $\delta = 0.5$

ペイロードサイズ：500 バイト

上記の条件において，Algorithm 1 により UAP を生成する．実験で用いるマルウェア検体は VirusShare [11] から入手可能な検体のうち文献 [2] の発表以前に流通している検体を用いるものとする．また，学習済み MalConv において高い確度でマルウェアと識別された検体から構成されたデータセットから UAP の生成を行う．

4.2 提案手法によって構築された汎用的敵対的摂動による識別への影響

Algorithm 1 によって生成された UAP の影響を明らかにするため，エラー率と信頼度減少率という 2 つの指標について測定を行う．ここで，$\mathrm{F}(X_i)$ をマルウェア検出器，$D = \{X_1, X_2, \ldots, X_{i-1}, X_i\}$ を入力検体群，V を汎用的摂動ベクトルとする．このとき，エラー率（Err）は式 (1) に定義される．同様に，信頼度減少率（Crr）は γ をしきい値とし式 (2) に定義される．

$$\mathrm{Err}(D, V) := \frac{1}{n}\sum_{i=1}^{n} 1_{F(X_i)\neq F(X_i+V)} \tag{1}$$

$$\mathrm{Crr}(D, V) := \frac{1}{n}\sum_{i=1}^{n} 1_{F(X_i+V)-F(X_i)\geq\gamma} \tag{2}$$

式 (1)，式 (2) によって算出された検証用データセットに対するエラー率，信頼度減少率を**表 1** に示す．このとき，信頼度減少率算出に用いるしきい値は $\gamma = 0.1$ とする．

MalConv への入力方法として，入力層からバイナリ値（離散値）を入力する方法と，MalConv の連続空間への入り口である埋め込み層以後の特徴量（連続値）を入力する方法を用いる．したがって入力バイナリを X_i とするとき，前者は $\mathrm{F}(X_i + V)$，後者は $\Pr(\mathrm{Em}(X_i) + V)$ を識別結果として用いる．また，生成に用いるマルウェア検体数ごとに測定を行い，比較対象として一様分布から生成されるランダムノイズについてもエラー率，信頼度減少率の測定を行う．

このとき，表 1 に示された実験結果から提案手法がランダムノイズと比べ，高い誤認を引き起こしていることを確認できる．

4.3 単一 UAP からの摂動複製手法

実験から，提案手法によりペイロード方式の摂動付与であったとしても UAP が計算可能であることが示された．しかし，1 つの UAP が計算可能であるとしても，計算した摂動が 1 度でも検出されてしまった場合，その後の検出においてはパターンマッチングなどの手法により，容易に防衛可能であることが考えられる．そこで，摂動 V から，V と同様に汎用的に誤認を引き起こす性質を持つ，V' を計算する手法を考える．

画像における UAP に対して，微細なゼロ平均ノイズを加えたとしても，UAP の持つパターンは保持され，多くの誤認能力を維持し続ける [7]．同様に MalConv では埋め込み層後の連続空間における摂動 v に微細なノイズを与えた場合，摂動の持つ誤認能力はやや低下しつつ引き継がれることが予想される．しかし，$\text{re-Em}(v)$ 関数により特徴量を離散値（バイナリ）へと再構成したとき，摂動 V，V' の各要素どうしは連続空間において近い距離であると扱われつつも，異なるバイナリ値となる．この性質を利用し，特徴量に対して正規分布から生成される乱数を加えることで，V' を生成する手法を考える．このとき，V' は式 (4) によって計算する．

$$\epsilon \sim N(0, \sigma^2 I) \tag{3}$$

$$V' = \text{re-EM}(\mathrm{Em}(V) + \epsilon) \tag{4}$$

各標準偏差 σ ごとの正規分布から生成されたノイズを用いて，式 (4) を適用することによって生成された V' のエ

表 2 標準偏差ごとの V' のエラー率およびバイナリ一致率
Table 2 Error rate and confidence reduce rate for V' calculated for each standard deviation.

$\sigma =$	0.01	0.05	0.1	0.5	1.0
エラー率	0.526	0.524	0.424	0.234	0.192
バイナリ一致率	0.874	0.482	0.204	0.010	0.002

ラー率およびバイナリ一致率を**表 2** に示す．バイナリ一致率とは V，V' の各要素どうしを比較し，一致している要素の割合である．また，V' のもととなる V は 1,000 検体を用いて生成したものを用いる．

表 2 から，乱数を用いて，1 つの UAP V から摂動 V' を多数生成可能なことが分かる．また，バイナリー一致率とエラー率はトレードオフの関係にあり，モデルごとに適切な乱数生成のパラメータを探索する必要がある．

5. まとめ

入力シーケンス長に制限がある実験環境の下であるが，提案手法が MalConv に対して有効な UAP を多数生成可能であることを示した．本論文における実験結果は MalConv [2] を含む，End to End 学習可能なマルウェア検出手法においても UAP が生成可能であることを示唆している．今後は，本論文を含めた敵対的サンプル生成手法を考慮し，Adversarial Training [6] をはじめとした検出器の堅牢性向上，適切な検出器の社会運用法の構築，長大なシーケンスを入力可能なモデルでの生成実験を行う必要がある．

参考文献

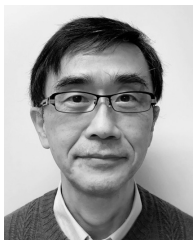
[1] Raff, E., Sylvester, J. and Nicholas, C.: Learning the pe header, malware detection with minimal domain knowledge, *Proc. 10th ACM Workshop on Artificial Intelligence and Security – AISec '17* (2017).
[2] Raff, E., Barker, J., Sylvester, J., et al.: Malware detection by eating a whole exe, *AAAI Workshops* (2018).
[3] Kancherla, K. and Mukkamala, S.: Image visualization based malware detection, *2013 IEEE Symposium on Computational Intelligence in Cyber Security* (*CICS*) (2013).
[4] Szegedy, C., Zaremba, W., Sutskever, I., et al.: Intriguing Properties of Neural Networks, *International Conference on Learning Representations* (2014).
[5] Kreuk, F., Barak, A., Aviv-Reuven, S., et al.: Deceiving end-to-end deep learning malware detectors using adversarial examples, arXiv:1802.04528 (2019).
[6] Goodfellow, I.J., Shlens, J. and Szegedy, C.: Explaining and harnessing adversarial examples, *International Conference on Learning Representations* (2015).
[7] Dezfooli, S.M., Fawzi, A., Fawzi, O., et al.: Universal Adversarial Perturbations, *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) (2017).
[8] Gao, H. and Oates, T.: Universal adversarial perturbation for text classification, arXiv:1910.04618 (2019).
[9] Hafemann L.G., Rony J., Abdoli, S., et al.: Universal adversarial audio perturbations, arXiv:1908.03173 (2020).
[10] Dauphin, Y.N., Fan, A., Auli, M., et al.: Language modeling with gated convolutional networks, *Proc. 34th International Conference on Machine Learning* (2017).
[11] Roberts, J.M.: VirusShare, available from ⟨https://virusshare.com/⟩ (accessed 2020-10-01).

宇野 貴士

1997 年生．2019 年信州大学工学部情報工学科卒業．同大学大学院修士課程在学中．深層学習におけるセキュリティ分野の研究に従事．

宮尾 秀俊

1989 年長岡技術科学大学電子機器工学課程卒業．1991 年長岡技術科学大学電気電子システム工学専攻修了．現在，信州大学工学部電子情報システム工学科准教授．博士（工学）．パターン認識・学習，HCI の研究に従事．電子情報通信学会，ACM 会各会員．



丸山 稔 （正会員）

1982 年東京大学工学部計数工学科卒業．同年三菱電機（株）入社．1990 年から 1991 年 MIT 人工知能研究所客員研究員．現在，信州大学工学部電子情報システム工学科教授．博士（工学）．コンピュータビジョン，機械学習等の研究に従事．電子情報通信学会，ACM，IEEE 各会員．