CNNによるマルウェア分類を改善するためのGANを用いたデータ拡張

CNN によるマルウェア分類を改善するための GAN を用いたデータ拡張

川畑　魁星 [a)]・油田　健太郎 [b)]・山場　久昭 [c)]・岡崎　直宣 [d)]

Data Expansion Using GAN to Improve Malware Classification by CNN

Kaisei KAWABATA, Kentaro ABURADA, Hisaaki YAMABA, Naonobu OKAZAKI

Abstract

In recent years, the spread of malware has become a threat to computer security. The existence of malware variants is a factor that has a significant impact on the increase in the number of malware discoveries. Research has been conducted to automatically and efficiently classify these variants of malware. With the development of deep learning, it is now used to classify subspecies of malware. A typical research is to convert malware into grayscale images and classify them using CNN (Convolutional neural network). In deep learning, a large amount of training data is used. However, when a new type of malware appears, it is difficult to collect a large amount of samples. In this research, we investigated whether it is possible to solve the problem of insufficient samples by generating training data for deep learning using GAN (Generative Adversarial Network) and extending the data. We conducted an experiment to see if the classification accuracy could be improved by expanding the data for training using GAN. We used datasets that consisted of 25 different malware families. It was confirmed that the classification accuracy was improved compared to that before the data expansion. From the results, it was found that the data expansion for malware classification using GAN was effective.

Keywords: Deep learning, Malware, Generative adversarial network, Data expansion

1. はじめに

マルウェアとは、不正かつ有害な動作を行う意図で作成された悪意のあるソフトウェアやコードの総称であり、コンピュータセキュリティにおける大きな脅威となっている。大手セキュリティベンダである McAfee から報告された 2019 年第 1 四半期のレポートでは、多数の新しいマルウェアのサンプルが発見された [1)]。マルウェアの発見数の増加に大きな影響を与えている要因として亜種の存在がある．亜種のマルウェアはツールを用いることで容易に作成できるため既存のマルウェアとはバイナリデータの異なるマルウェアが大量に作成される。亜種のマルウェアが増加する中で専門的な解析を用いずに、効率よく自動的に分類を行うことで対象のマルウェアに応じた対応を迅速に行うことができる。

マルウェアの脅威を防ぐ対策として行われるマルウェア解析は静的解析手法と動的解析手法に大別される。静的解析手法はマルウェアをリバースエンジニアリング等の技術によって解析を行う手法である。マルウェアのバイナリコードを解析することで解析対象のマルウェアがどのような機能をもっているか詳細に把握することができる。しかし、静的解析手法を行うには高度な専門知識が必要であり、複雑なマルウェアである場合、時間的コストが必要となる。一方、動的解析手法はマルウェアを実際に動作させ、その挙動を監視することで解析結果を取得する解析手法である。難読化されたマルウェアでも解析できるため静的解析手法と比較し解析に要するコストが低いが、サンドボックスなどの特殊な環境が必要となる。

そこで、静的解析手法や動的解析手法を使用せず深層学習を用いてマルウェアの亜種の分類を自動的に行う手法が存在する。深層学習の発達に伴い畳み込みニューラルネットワーク（CNN：Convolutional neural network）等を用いた画像分類が急速に発展を遂げている。マルウェアのバイナリデータをグレースケールの画像に変換し、CNN に適用することで既知のどのマルウェアの亜種であるか分類を行う。これによりマルウェアの分類を多クラスの画像分類問題に置き換えることができ、マルウェアに関する専門的な高度な知識を必要とせず分類できる。深層学習による画像分類では一般的にラベル付けされた大量の学習用のデータが必要である。しかし、新しいタイプのマルウェアが出現した場合など、サンプルを集めラベル付けを行い学習用のデータを大量に用意するのは非常に困難である。

サンプルを大量に用意できない場合、従来の方法ではオリジナルのデータに対して図 1 のような拡大、回転、クロップなどの幾何変換を施すことによってデータ拡張を行い、サンプルが少ない問題を解決をしていた [2)]。しかし、マルウェアをグレースケールに変換した画像は一般的な画像と異なっており従来の方法でのデータ拡張は不向きである。そのため深層学習を用いた画像の生成方法を用いてデータ拡張を行う。具体的に

[a)] 工学専攻機械・情報系コース大学院生
[b)] 情報システム工学科准教授
[c)] 情報システム工学科助教
[d)] 情報システム工学科教授

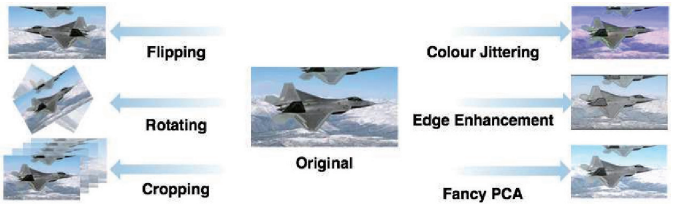

図 1. 幾何変換によるデータ拡張（[2] より引用）

表 1. マルウェアのファイルサイズと画像の幅の対応

ファイルサイズ	画像の幅
<10KB	32
10KB ～ 30KB	64
30KB ～ 60KB	128
60KB ～ 100KB	256
100KB ～ 200KB	384
200KB ～ 500KB	512
500KB ～ 1000KB	768
1000KB≤	1024

は、敵対的生成ネットワーク（GAN：Generative Adversarial Network）を用いてオリジナルのマルウェア画像を基に擬似的なマルウェア画像を生成しデータ拡張を行う。

2. 関連研究

マルウェアのバイナリデータをグレースケールの画像に変換し自動的に分類する手法は、Nataraj らによって提案されている [3]。マルウェアの画像化を行い分類する具体的な手順は以下の通りである。

1. マルウェアのバイナリデータを 1 バイトずつ読み込み、0～255 の整数列へと変換する。
2. マルウェアのファイルサイズに基づいて、表 1 のように幅を決め、整数列をグレースケール（0:黒、255:白）の画像に変換する。
3. 得られた画像を 64 × 64 の解像度に伸縮する。
4. 伸縮させたグレースケールの画像から、GIST と呼ばれる画像のテクスチャ特徴を計算するための手法を用いて 320 次元の特徴量を抽出し、特徴ベクトルとする。

以上の手順により、既知のマルウェアから特徴ベクトルを抽出し、その特徴ベクトルを k 近傍法（k=3）の分類器に学習させ新たなマルウェアを分類する。25 種類のマルウェアファミリーを持つ 9,458 サンプルのマルウェアのデータセットを用いて実験を行い 98 %の分類精度を実現したと報告されている。

Nataraj らは、[4] において、既存手法より高い精度で分類することができたと報告している。特に、難読化されている、パックされたマルウェアに対しても、画像化を行い分類する手法は有効であったと述べている。

3. 関連技術

3.1 畳み込みニューラルネットワーク

文献 [3] はマルウェアをグレースケールの画像に変換し、既知のマルウェアの画像から特徴ベクトルを抽出する。抽出した特徴ベクトルを機械学習の手法である k 近傍法の分類器に学習させマルウェアを各ファミリに分類した。本研究では、文献 [3] の手法でマルウェアをグレースケールの画像に変換し CNN を用いてマルウェアファミリの分類を行う。

近年、深層学習は物体認識や自然言語処理、音声認識などの幅広い分野で使用されている。深層学習の手法は数多く提案されており、画像認識の分野においても様々なアプローチが検討されてきたが、現在最も成功を収めているのが CNN である。CNN は、多層のニューラルネットワークからなり、一般的なニューラルネットワークに用いられる全結合層に加え、画像の局所的な特徴抽出を担う畳み込み層と、局所ごとに特徴をまとめ小さな位置変化に対して頑健性を高めるプーリング層が含まれていることが特徴である。

畳み込み層は入力データに対し、フィルターを適用し畳み込み演算を行う。畳み込み演算を行うことで局所的な入力パッチから特徴量を抽出し、表現のモジュール化とデータの効率化を可能にする。

3.2 敵対的生成ネットワーク

3.2.1 GAN

新しいタイプのマルウェアが出現した場合など、深層学習手法に必要なラベル付きの学習用データを大量に用意することが困難である。そのため、深層学習手法に必要である学習用データを大量に生成するため本研究では GAN を用いて画像の生成を行う。

近年、学習データと類似のデータを生成する確率的な深層生成モデルとして、敵対的生成ネットワーク（GAN：Generative Adversarial Network）[5] や変分自己符号化器（VAE：Variational Autoencoder）[6] が盛んに研究されている。GAN は、2014 年に Ian J. Goodfellow らによって提案された教師なしの生成モデルである。

GAN の目的は、真のデータの分布を $P_r(x)$、生成データの分布を $P_g(x)$ としたとき、$P_r(x)$ と一致するような $P_g(x)$ を得ることである。これを実現するために生成器（Generator）と識別器（Discriminator）の 2 つのネットワークを用いる。生成器 G は乱数 z～$P_z(x)$ からデータ空間 $x = G(z)$ への写像を行う。識別器 D はデータ x が $P_r(x)$ からサンプリングされたものであれば確率 $p = D(x) \in [0, 1]$ を付与し、$P_g(x)$ からサンプリングされたものであれば確率 $1 - p$ を付与する。D と G は以下の Min-Max の目的関数で最適化が行われる。

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log 1 - D(G(z))] \quad (1)$$

上式では、G は目的関数の最小化を行うことで、D が真のデータと見分けのつかない、つまり D を「騙せる」ようなデータを生成できるようにする。一方、D は目的関数の最大化を行うことで、真のデータと生成データの識別境界を見つけ、G

に「騙されない」ようにする。このように、G と D が敵対的な関係を持ち、競争しながら最適化される点が GAN の特徴である。

VAE と異なり $P_g(x)$ は最適な条件下では $P_r(x)$ に漸近することが示されており、緻密なデータ生成が可能である。短所として、Min-Max 最適化を行うため、学習が安定しない点が挙げられる。

3.2.2　DCGAN

上記で述べた GAN の問題点である学習が安定しない点を解決したのが深層畳み込み GAN（DCGAN：Deep Convolutional GAN）[7] である。DCGAN では GAN を安定的に学習するためネットワークの設計を改良している。要点は以下の 4 つにまとめられる。

1. プーリング層を、D ではすべて畳み込み層に、G では逆畳み込み層に置き換える。
2. G では活性化関数として出力層以外は Rectified Linear Unit（ReLU）を用い、出力層には Tanh を用いる。
3. D では活性化関数に Leaky ReLU を用いる。
4. Batch Normalization を用いる。

DCGAN による学習の安定化は、精密な画像生成を可能にするのに加え、表現力の高い潜在空間の学習も可能にする。本研究では、この DCGAN を用いてマルウェア画像の生成を行う。

4.　提案手法

本研究は、GAN よりも学習が安定し、精密な画像を生成することが可能な DCGAN を用いてマルウェア画像を生成し、CNN でマルウェアの亜種を分類するのに必要な学習データのデータ拡張を行う手法を提案する。深層学習手法でマルウェアの亜種を分類するにあたって、データセットとして大量の学習データが必要である。しかし、新しく出現したマルウェアなど、分類対象によっては十分な数のサンプルを集めることが困難な場合がある。そこで、GAN を用いて学習データを自動的に生成することでマルウェア画像のデータ拡張を行う。本章では、データ拡張に用いるマルウェア画像の生成方法について述べる。

4.1　データ拡張

4.1.1　幾何変換

従来の手法ではオリジナルの画像に拡大・縮小、ランダム回転、ランダムズーム、反転、シフト、シアー変換などを行いデータの拡張を行っていた。しかし、マルウェアの画像は一般的な画像とは異なるため、幾何変換の処理を施すとオリジナルの画像とかけ離れた画像となってしまい分類精度の低下が予想される。本研究では、分類精度に影響が一番出ないであろうと考えられる水平反転を用いてデータ拡張を行う。各クラスの画像に画像処理を行い、訓練データの枚数を 2 倍にする。GAN で生成された画像に多様性を出すために訓練データに、オリジナルの画像と水平反転の処理を施した画像を使用する。

[1] オリジナル　[2] 水平反転　[3]DCGAN

図 2. マルウェアファミリ C2LOP.gen!g の画像

[1] オリジナル　[2] 水平反転　[3]DCGAN

図 3. マルウェアファミリ Fakerean の画像

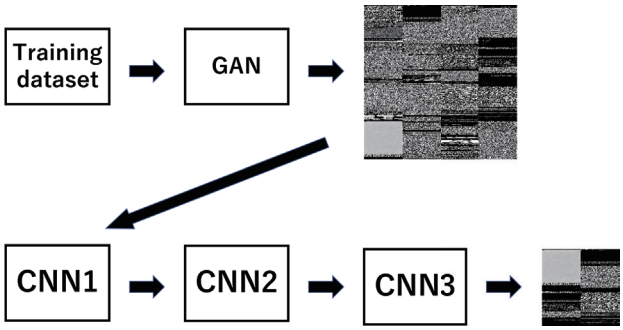

図 4. 生成画像の評価の流れ

4.1.2　DCGAN を用いた画像生成

ネットワークの構造は生成器が 4 層の逆畳み込み層と全結合層、識別器が 4 層の畳み込み層と全結合層からなる。100 次元の一様乱数を入力とし、64 × 64pixel の画像を出力として得る。マルウェアファミリそれぞれのネットワークを作成し学習させる。

2 つのマルウェアファミリのオリジナルの画像を水平反転させた画像と DCGAN によって生成された画像を図 2、3 に示す。

4.2　生成画像の評価

DCGAN で画像を生成した際に、学習に使用したデータ画像からかけ離れた画像が生成されることが起きる。それらの画像を用いて CNN で学習を行うと分類精度の低下を引き起こす可能性がある。そこでデータセットで事前に学習済みの CNN を用いて画像の選別を行う。CNN がラベルとクラスが同じであると分類した画像のみを実験に使用する。生成画像の評価の流れを図 4 に示す。

5.　分類精度の評価実験

DCGAN によって生成したマルウェア画像を CNN の訓練データに加えることでデータ拡張を行えるか検証する。マルウェア画像の生成には DCGAN を用い、データ拡張を行わない場合、幾何変換のみを用いてデータ拡張を行った場合とど

表 2. Malimg Datasets の内訳

No	ファミリ名	タイプ	検体数
0	Adialer.C	Dialer	122
1	Agent.FYI	Backdoor	116
2	Allaple.A	Worm	2949
3	Allaple.L	Worm	1591
4	Alueron.gen!J	Trojan	198
5	Autorun.K	Worm:AutoIT	106
6	C2LOP.gen!g	Trojan	200
7	C2LOP.P	Trojan	146
8	Dialplatform.B	Dialer	177
9	Dontovo.A	Trojan Downloader	162
10	Fakerean	Rogue	381
11	Instantaccess	Dialer	431
12	Lolyda.AA1	PWS	213
13	Lolyda.AA2	PWS	184
14	Lolyda.AA3	PWS	123
15	Lolyda.AT	PWS	159
16	Malex.gen!J	Trojan	136
17	Obfuscator.AD	Trojan Downloader	142
18	Rbot!gen	Backdoor	158
19	Skintrim.N	Trojan	80
20	Swizzor.gen!E	Trojan Downloader	128
21	Swizzor.gen!I	Trojan Downloader	132
22	VB.AT	Worm	408
23	Wintrim.BX	Trojan Downloader	97
24	Yuner.A	Worm	800

の程度分類精度に違いがあるかを比較する。

5.1 データセット

本研究ではマルウェアデータセットとして、Nataraj らによって作成された Malimg dataset[3] を用いる。Malimg dataset は 25 種類のマルウェアファミリで構成され、9,339 検体のマルウェア画像があり各クラスのサンプル数は異なる。表 2 に Malimg dataset の内訳を示す。

5.2 実験手順と実験結果

今回は少数のデータセットに対する提案手法の効果の検証を行う。本研究では、5 分割交差検証法によって分類精度を算出した。各クラスからランダムに画像を 30 枚抽出しテストデータとして用いる。テストデータを抽出した後、各クラスから訓練データを抽出する。各クラスの残っている画像が 170 枚より少ない場合、残りすべてを訓練データとして使用する。残っている画像が 170 枚を超える場合、170 枚をランダムに抽出する。

訓練データに Malimg dataset を用いて、25 クラスそれぞれ DAGAN のネットワークを作成する。DCGAN で画像の生成を行い、Malimg dataset で事前に学習済みの CNN を 3 つ用いて画像の評価を行う。CNN の訓練データに評価した画像を加え学習させ、テストデータに対しての分類精度を算出する。加える画像の枚数による分類精度の変化を検証するため各クラス 100 枚ずつ増やし、最大 500 枚まで訓練データに加えた。手法による分類精度の違いを表 3 に示す。

表 3. 分類精度

手法	分類精度
CNN のみ	94.9 %
水平反転によるデータ拡張	96.1 %
提案手法によるデータ拡張 (100 枚)	96.2 %
提案手法によるデータ拡張 (200 枚)	96.2 %
提案手法によるデータ拡張 (300 枚)	96.4 %
提案手法によるデータ拡張 (400 枚)	96.5 %
提案手法によるデータ拡張 (500 枚)	96.9 %

表 4. ファミリ別の分類精度

ファミリ名	データ拡張前	データ拡張後
Allaple.A	95.0 %	97.3 %
C2LOP.gen!g	95.0 %	98.6 %
C2LOP.P	88.3 %	92.0 %
Swizzor.gen!E	58.3 %	66.6 %
Swizzor.gen!I	60.0 %	67.3 %

ベースラインとなる、Malimg dataset を学習させデータ拡張を行っていない CNN の分類精度が 94.9 %だったのに対し、一番分類精度の高い提案手法を用いて各クラス 500 枚加えたデータ拡張では 96.9 %と 2 %分類精度が向上した。幾何変換だけを行いデータ拡張を行った CNN の分類精度が 96.1 %だったことから、マルウェア画像に対し、GAN を用いたデータ拡張は効果があることが確認できた。

分類が困難であったマルウェアファミリのファミリ別の分類精度の比較を表 4 に示す。データ拡張を行うことで Swizzor.gen!E、Swizzor.gen!I などの他のマルウェアファミリに比べ分類精度が低いファミリの分類精度の改善が確認できた。

6. まとめ

本研究では、GAN を用いてマルウェアの画像のデータ拡張を行う手法を提案した。この提案により少数のデータセットに対するマルウェアの分類精度が低くなる問題を解決することができる。結果として、縮小した Malimg dataset に対して最大で 2 %の分類精度の上昇が確かめられた。幾何変換だけでデータ拡張を行った場合よりも 0.8 %の分類精度の上昇が確かめられた。GAN を用いてデータ拡張を行うことで新しく出現したマルウェアで、サンプルの収集が困難である場合などにも対応することが可能となる。学習用データは正しくラベル付けされている必要があるが、大量の学習用データに正しくラベル付けを行う作業はコストがかかる。GAN を用いて生成されたデータはラベル付けを行う必要がないためコストを削減することが可能である。今後の課題としてマルウェアの検体がどの程度あれば GAN でデータ拡張に使用できるクオリティのマルウェア画像を生成可能であるか調査を行う必要がある。

参考文献

1) McAfee Labs: 脅威レポート：2019 年 8 月,
https://www.mcafee.com/enterprise/en-us/

assets/reports/rp-quarterly-threats-aug-2019.pdf

2) Luke Taylor, Geoff Nitschke: Improving Deep Learning using Generic Data Augmentation, IEEE Symposium Series on Computational Intelligence (SSCI), 2018.

3) Nataraj, L. et al.: Malware images: visualization and automatic classification, VizSec, p. 4, 2011.

4) Nataraj, L. et al.: A comparative assessment of malware classification using binary texture analysis and dynamic analysis, AISec, pp. 21–30, 2011.

5) Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets., Advances in neural information processing systems , pp. 2672-2680, 2014.

6) Kingma, D. P., Welling, M.: Auto-encoding variational bayes., arXiv preprint, arXiv:1312.6114, 2013.

7) A. Radford, L. Metz and S. Chintala: Unsupervised representation learning with deep convolutional generative adversarial networks, Proc. ICLR, 2016.