7D-02

# VQ-VAE を用いたマルウェア検知

ダオ　ヴァン　トゥアン†　　佐藤　浩†　久保　正男†

防衛大学校情報工学科†

## 1．　はじめに

マルウェアの増加や複雑化は，サイバーセキュリティ上の最も深刻な脅威の一つとなっている．サイバーセキュリティ業界は，常にマルウェアを監視し，様々な方法でその検知や対処に取り組んでいるが，サイバー攻撃者の攻撃が止む兆候はない．従来の静的・動的解析手法は，新しいマルウェアの特定には効果がなく，メモリと時間の点で高い負荷がかかっている．また，手動で作成した特徴量に基づいて分類器を学習する一般的な機械学習アプローチも，こうした回避的な手法に対しては十分な威力を発揮せず，特徴量エンジニアリングのために多くの労力を必要とする．

機械学習は，優れたアルゴリズムの開発により，各分野で遭遇する問題の多くを解決する程強力になっている．マルウェア検知のための機械学習ベースの方法の一つに，ビジョンベースのアプローチがある．プログラムバイナリ（マルウェア／良性）を画像として可視化することで，プログラムコードの各セクションを明確に認識することができる．マルウェアの作者は，オリジナルのコードの一部に修正を加え，新しい亜種を作り出す．画像化は全体の構造を保持しつつ，小さな変化も捉えることができる．元のコードからの逸脱は，パターンのバリエーションによって識別される．攻撃者は難読化技術を使用するため，同じマルウェアの亜種では，同じ場所に同一のコードが位置しない可能性があるが，全体としてある程度類似したコードとデータの順序は維持されている．そのため，機械学習の手法である畳み込みニューラルネットワーク（CNN）は，マルウェアのバイナリから亜種間に共通する機能を抽出できる．

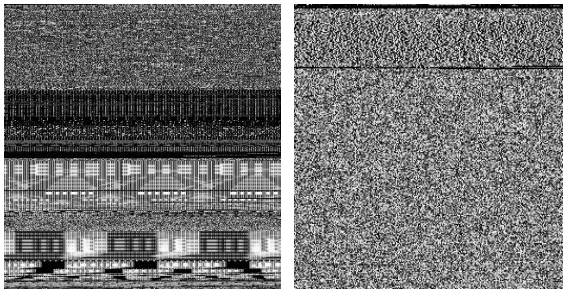
既存研究では，複雑なニューラルネットを構築することが多いが，その実行には高性能なコンピュータと長い処理時間が必要である．本究では VQ-VAE[1]を軽量化したモデルにより，画像化されたデータセットに対しマルウェアの分類を行った．本手法はシンプルかつ軽量なモデルであり，短い時間で高い検知精度を得ただけでなく，難読化されたマルウェアも識別することもできた．

Image-based Malware Detection Using VQ-VAE

† Graduate School of Science and Engineering, National Defense Academy

## 2.　プログラムバイナリファイルの画像化

プログラムバイナリファイルを画像として視覚化するには，各バイトを画像の 1 ピクセルとして解釈する必要がある．バイナリファイルの各 16 進数のペアを 1 つの 10 進数とし，画像のピクセル値として扱う．その値をグレースケールと対応付けて画像に変換する（図１）



良性ファイル　　マルウェア

図１　プログラムバイナリファイルを画像化したサンプル

## 3.　提案手法

画像から特徴を抽出するモデルは様々あるが，その中で，畳み込みニューラルネットが一般的に使われている．最後の全結合層で全ての特徴をまとめ Sigmoid 活関数及び binary cross entropy といった損失関数で検知が行われる．これらの手法は教師あり学習である．精度改善のためには，大きくのデータが必要である．一方，教師なし学習では，少ないデータでも対応できるが，如何に良い特徴ベクトル空間を生成するかが課題となる．VQ-VAE はベクトル量子化（VQ）により，学習が簡単で，大きな分散に悩まされることもなく，強力なデコーダを持つ多くの VAE モデルで問題となる，潜在的な情報が無視されてしまうことによる「事後崩壊」問題を回避することができる．VQ-VAE は潜在空間を有効に利用できるため，通常データ空間の多くの次元にまたがる重要な特徴をうまくモデル化することができる．

VQ-VAE のアーキテクチャーを図２に示す．モデルは入力 x を受け取り，それをエンコーダに通して出力 $z_e(x)$を生成する．次に，離散潜在変数 z は，共有された埋め込み空間 e を用いて，最近傍探索により計算される．復号器への入力は，対応する埋め込みベクトル$e_k$となる．



Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

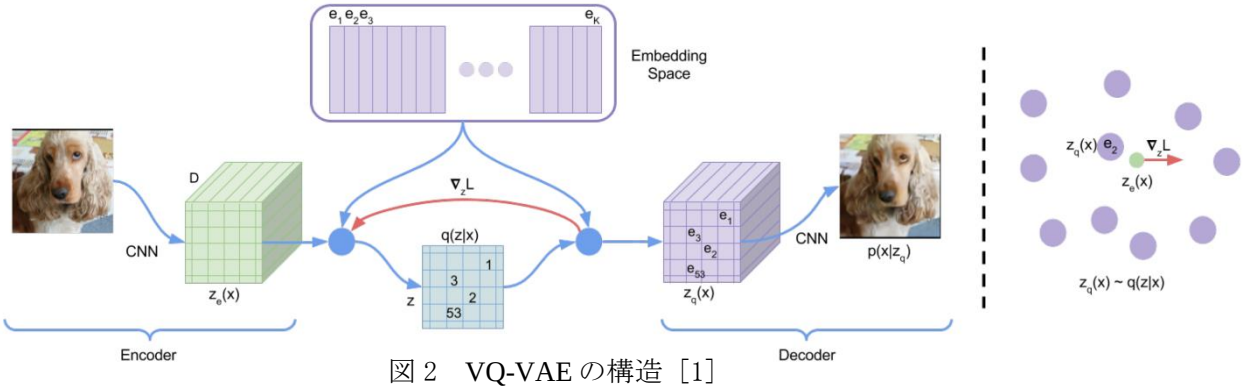



図２　VQ-VAE の構造［1］

本研究では，訓練時に VQ-VAE を用いて，画像化したマルウェアよりマルウェアの特徴潜在空間を生成し，機械学習の古典的な分類器で分類を行う．

## 4. 実験

本研究では Virustotal [2]より 1000 のマルウェアサンプルを収集し，良性サンプルについては，1000 の Windows のシステムディレクトリ（…\Windows\System32）にある実行ファイルのうち，様々な機能や特徴を持つファイルを選択した．表１より，Random Forest（RF）分類機は 97%と最も高い Accuracy が得られた．一方 Nearest Centroid はわずか 0.06 秒の学習時間であるが，RF と比べ，4.28%の大きな Accuracy の差がある．表２は，CNN モデルの代表的な巨大（VGG19）および軽量（MobileNet）モデルとの比較結果である．VGG19 は大量のパラメータにも関わらず低い精度である．VQ-VAE は MobileNet と比べ Accuracy 差が 2%であるが，わずか 1/60 のパラメータ数であるため，より軽量なデバイスとして幅広く利用できる．

表１　分類器の違いによる性能の違い

| 分類機 | Accuracy (%) | Time(s) |
|---|---|---|
| Decision Tree | 97.18 | 2.40 |
| kNN | 98.23 | 0.32 |
| Naïve Bayes | 94.75 | 0.28 |
| Nearest Centroid | 92.72 | **0.06** |
| **Random Forest** | **97.00** | 3.41 |
| SVM | 95.50 | 1.80 |

表２　他のモデルとの性能比較

| モデル | Accuracy (%) | #Parameters(M) |
|---|---|---|
| VGG19 [3] | 52.25 | 45.21 |
| **MobileNet [4]** | **99.00** | 3.23 |
| VQ-VAE (Random Forest) | 97.00 | **0.05** |

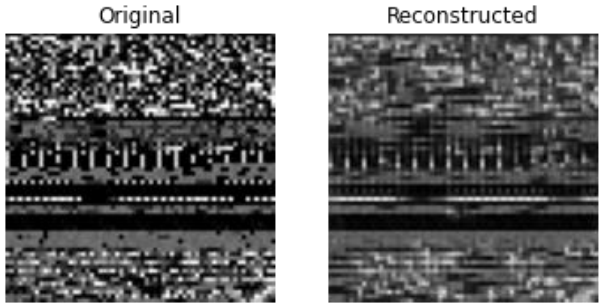



図３　元画像と VQ-VAE で再現した画像

図３の左は難読化されたマルウェアの画像化であり，VQ-VAE を用いて学習し，再現した画像は右である．左右の図を比べると，ある程度の再現ができたと言える．これによって，良い潜在空間の生成ができることが分かった．

## 5. まとめ

本研究では，教師なしの VQ-VAE を用いたマルウェア検知を行った．VQ-VAE は，深いネットワークを使った既存研究と比べ，わずかなパラメータで CNN モデルと匹敵する結果を示した．VQ-VAE は低パワーの IoT への採用することが期待される．今後は，さらなる精度の改善のため，VQ-VAE の構造など複雑さに関わる部分を改良することが必要である．

## 参考文献

[1] V. D. Oord, O. Vinyals, and K. Kavukcuoglu, “Neural Discrete Representation Learning,” in Advances in neural information processing systems, 2017.
[2] https://www.virustotal.com
[3] K. Simonyan, and A. Zisserman. “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
[4] A. G. Howard, M. Zhu, B. Chen, and D. Kalenichenko, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” arXiv preprint arXiv:1704.04861, 2017.



Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.