VAEを用いた ネットワークトラフィックの異常検出

中前 諒哉[1,a)]　青木 茂樹[1,b)]　宮本 貴朗[1]

概要：一般的な深層学習手法を IDS （Intrusion Detection System）に応用する場合，教師データとして必要な，多岐にわたる異常データの収集が困難であることが課題となっている．そこで本稿では，深層学習の中でも教師データを必要としない VAE（Variational AutoEncoder）を応用した IDS の構築手法を提案する．VAE は学習した画像と類似する画像は再構成できるが，学習していない画像に類似する画像は再構成できないため，この性質を利用してネットワークトラフィックの異常を検出する．まず，学習期間のトラフィックデータを画像に変換して VAE で学習する．次に，別の期間に収集したトラフィックデータを学習時と同様に画像に変換し，変換した画像に類似する画像を VAE で再構成する．そして，入力した画像と再構成した画像の差分を基に異常度を算出し，算出した異常度により正常な通信と異常な通信を識別する．実験では，MWS2018 データセットと CICIDS2017 データセットを用いて本手法の有効性を確認した．

キーワード：ネットワーク異常検出，深層学習，VAE

Anomaly Detection of Network Traffic using Variational AutoEncoder

Ryoya Nakamae[1,a)]　Shigeki Aoki[1,b)]　Takao Miyamoto[1]

Abstract: In general, deep learning method requires teacher data. However, when applying deep learning methods to Intrusion Detection System, it is difficult to collect various anomaly data. In this paper, we propose a method for constructing IDS using Variational AutoEncoder (VAE) which does not require teacher data. VAE can reconstruct images similar to learned images, but cannot reconstruct images similar to unlearned images. We use this property to detect anomaly in network traffic. First, we convert traffic data during learning period into images and learn the images by VAE. Second, we convert traffic data collected in another period into images in the same way as during learning, and use VAE to reconstruct images similar to the converted images. Anomaly scores are calculated based on the difference between input images and reconstructed images. Using calculated anomaly scores, we detect anomaly communication. We confirmed the effectiveness using MWS2018 dataset and CICIDS2017 dataset.

Keywords: Anomaly Detection of Network, Deep Learning, Variational AutoEncoder

1. はじめに

近年，標的型攻撃等のネットワーク犯罪の増加に伴い，ネットワーク上の不正なトラフィックを検出する侵入検知システム（IDS：Intrusion Detection System）の研究が盛

[1] 大阪府立大学大学院人間社会システム科学研究科
Graduate School of Humanities and Sustainable System Sciences, Osaka Prefecture University
[a)] saa01183@edu.osakafu-u.ac.jp
[b)] aoki@kis.osakafu-u.ac.jp

んに行われている．IDS はシグネチャ型とアノマリ型の 2 種類に分けることができる．代表的なシグネチャ型 IDS として，Snort [1] や Suricata [2]，TheBro [3] 等が挙げられる．シグネチャ型 IDS は異常を定義したパターンファイルに基づいて異常を検出する方式である．この方式には，パターンファイルに定義されていない攻撃は亜種を含めて検出できない欠点が存在する．一方，代表的なアノマリ型 IDS として文献 [4–6] の手法が挙げられる．アノマリ型 IDS は正常な通信のみを含むデータから正常状態を定義し，正常から外れた状態を異常として検出する方式である．この方式では，シグネチャ型 IDS の問題点であった未知の異常を検出できる．しかし，シグネチャ型 IDS に比べ，誤検知が多く発生することが欠点として挙げられる．

本研究では，アノマリ型 IDS の精度向上のために，画像処理分野において高い識別性能を示している深層学習に注目する．一般的に用いられる深層学習手法では学習時に教師データが必要となる．しかし，ネットワークの異常検出に応用する場合，正しく正常と異常のラベルがついたデータを取得することは難しい．そこで，深層学習の中でも教師データを必要としないオートエンコーダ（AE : AutoEncoder）[7] に注目する．AE は Encoder と Decoder の 2 つのニューラルネットワークを用いて，入力データに一致するデータを出力することを目的とするモデルである．文献 [8] では，文献 [7] の AE の潜在変数部分に確率分布を導入することで更に精巧な画像を再構成することが可能な VAE（Variational AutoEncoder）を提案している．VAE には学習した画像と類似する画像は再構成できるが，学習していない画像に類似する画像については再構成できない性質が存在する．文献 [9] では，この性質を利用して信号波形の画像データから既知の画像データか未知の画像データかを識別する手法を提案している．本稿では文献 [9] を応用し，既知のトラフィックデータの画像を学習した VAE により，正常なトラフィックデータの画像と未知の異常なトラフィックデータの画像を識別する手法を提案する．まず，既知の通信のトラフィックデータを画像に変換して VAE で学習する．その後，テスト用のトラフィックデータを学習時と同様に画像へと変換し，変換した画像に類似する画像を VAE で再構成する．そして，入力した画像と再構成した画像との差分を基に異常度を算出し，算出した異常度により正常な通信と異常な通信を識別する．実験では，MWS 2018 データセットの BOS データセット [10] と CICIDS 2017 データセット [11] を用いて本手法の有効性を確認した．以下，2 節で関連研究，3 節で提案手法，4 節で実験と考察，5 節でまとめについて述べる．

2. 関連研究

2.1 ネットワークの異常検知に関する関連研究

本研究に関連する従来研究として，アノマリ型 IDS に関する文献 [4–6] の概要を説明する．文献 [4] では，パケットのエントロピーに基づく異常検出手法が提案されている．この手法ではまず，単位パケット数あたりの IP アドレスやポート番号などの出現回数を計測する．次に，出現回数を基に特徴量の出現確率を求め，求めた出現確率からエントロピーを算出する. その後，エントロピーの時系列変化に注目した EMMM 法（Entropy-based Multidimensional Mahalanobis distance Method）により，エントロピーが大きく変化する場合を攻撃などが含まれている異常状態として検出している．文献 [5] では，複数の特徴量の組み合わせによる異常検出手法を提案している．この手法では，異常をトラフィック量の異常，通信手順の異常，通信内容の異常の 3 種類に分け，単位時間あたりのトラフィック量を数値化した特徴量，フロー毎のフラグの出現回数等を数値化した特徴量，フロー内のパケットのペイロードのパターンの傾向を数値化した特徴量を学習用データからそれぞれ抽出する．そして，新たなデータでこれらの特徴量を抽出し，学習用データの値と閾値以上離れている特徴量が存在する場合に異常であると判断する. 文献 [6] では，ネットワークのトラフィックは複数の正常状態で表されると考え，複数の正常状態を定義し，各状態との違いから異常を検出する手法を提案している．この手法では，異常を含まないデータから単位時間当たりの ICMP や TCP パケット数等を計測してクラスタリングする．メンバが少ないクラスタは削除し全てのクラスタにおいて閾値以上のメンバ数となるまでクラスタリングを繰り返す．クラスタリング結果を正常状態として定義し，新たに観測されたデータから同様の特徴を抽出し，正常クラスタとの距離が閾値以上であるか否かで異常の判別を行っている．

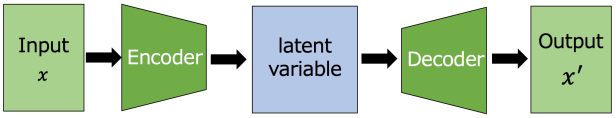

図 1　オートエンコーダーの構成
Fig. 1　Constitution of AutoEncoder.

2.2 AE に関する関連研究

オートエンコーダ（AE : AutoEncoder）に関する従来研究として，文献 [7–9] について述べる．文献 [7] では，入力データを学習し，入力データに精巧に類似するデータを再構成できる AE を提案している．AE の構成を図 **1** に示す．AE は深層学習の中でも，教師データを必要としない手法である．AE は 2 つの構成要素 Encoder と Decoder から成る Deep Newral Network の一種である．Encoder は入力したサンプル $\boldsymbol{x}$ を低次元の中間変数 $\boldsymbol{z}$ に写像する．Decoder は中間変数 $\boldsymbol{z}$ を受け取り，元のサンプル $\boldsymbol{x}$ を再構成するように $\boldsymbol{x'}$ を出力する．AE はデータの圧縮と再構

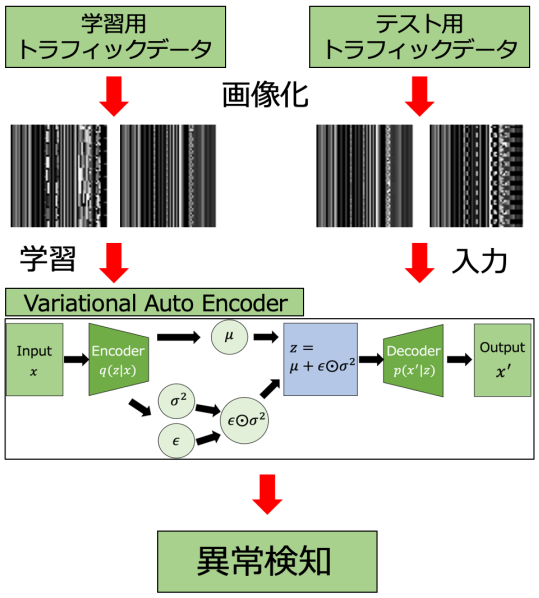

図 2　提案手法の概要

Fig. 2　Overview of proposed method.

成を行うモデルである．一方，AE に確率分布の理論を導入したモデルが変分オートエンコーダ（VAE：Variational AutoEncoder）[8] である．文献 [8] の手法では，確率分布の導入により学習を安定させ，更に精巧な画像を再構成できる．AE が元のサンプル $\boldsymbol{x}$ に類似した $\boldsymbol{x'}$ を再構成できるように，VAE も学習したサンプル $\boldsymbol{x}$ と類似する画像 $\boldsymbol{x'}$ を再構成できるが，学習していないサンプル $\boldsymbol{x}$ を再構成した $\boldsymbol{x'}$ は $\boldsymbol{x}$ に類似しない．文献 [9] では，この性質を応用して信号波形の画像データの異常検知を行っている．まず，学習用データとして正常な信号波形の画像 $\boldsymbol{x}$ を VAE で学習する．その後，テスト用の信号波形の画像 $\boldsymbol{y}$ を学習済みの VAE に入力し，類似する画像 $\boldsymbol{y'}$ を VAE で再構成する．入力した画像 $\boldsymbol{y}$ が学習した画像 $\boldsymbol{x}$ に類似する場合，確率分布を正確に求めることができ，入力した画像 $\boldsymbol{y}$ と再構成した画像 $\boldsymbol{y'}$ との差分が小さくなる．しかし，学習した画像 $\boldsymbol{x}$ に類似しない未知の信号波形の画像 $\boldsymbol{y}$ の場合，確率分布を正確に求めることができず，入力した画像 $\boldsymbol{y}$ と再構成した画像 $\boldsymbol{y'}$ との差分が大きくなる．この性質を利用して求めた差分を基に，異常度を算出することで異常な信号波形の画像を検出している．

3.　提案手法

提案手法の概要を図 **2** に示す．本手法は学習と異常検知の 2 つのプロセスに分かれている．まず，学習プロセスでは，学習期間中のトラフィックデータを画像に変換し，学習用データとして VAE で学習する．異常検出プロセスでは，学習とは別に取得したトラフィックデータを学習プロセスと同様に画像に変換し，テスト用データとする．その後，学習済みの VAE により正常な通信と異常な通信を識別する．

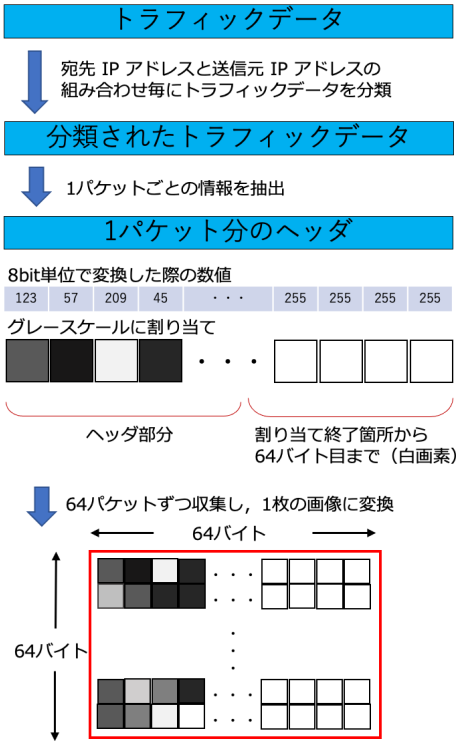

図 3　画像への変換手法の概要

Fig. 3　Overview of image conversion method.

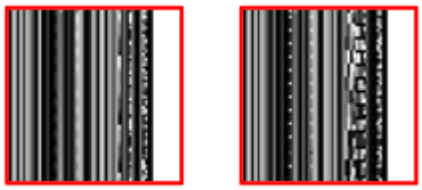

図 4　画像への変換例（左：正常通信 右：異常を含む通信）

Fig. 4　Examples of conversion image（left：normal communication right：anomaly communication）.

3.1　トラフィックデータの画像変換

本研究では，特定のホストから別の特定のホストに対するパケットに注目して異常を検出する．そこで，宛先 IP アドレスと送信元 IP アドレスの組み合わせ毎に観測されたパケットを画像に変換する．特定のホストから特定のホストに対する通信のみを抽出することによって，ブルートフォース攻撃や DoS 攻撃，クロスサイトスクリプティングなどの様々な攻撃の特徴を顕著に表現できると考えられる．また，近年のトラフィックの特徴として SSL 通信が多用されていることが挙げられる．暗号化された通信でペイロードの特徴を用いると正しく学習できない可能性が高いため，ここではヘッダ部分のみを用いて画像に変換することとする．画像への変換の概要を図 **3** に示す．あるネットワークで送受信されるパケットを pcap 形式でキャプチャしてトラフィックデータとする．宛先 IP アドレス

と送信元 IP アドレスの組み合わせ毎にキャプチャしたトラフィックデータを分類する．そして，抽出したパケットを 1 パケットずつ読み込み，パケットのヘッダ部分を 8 bit 単位で 0∼255 の数値に変換する．変換した数値を 1 画素のグレースケールの値に割り当てる．ここで，VAE で学習する際に画素が 8 の倍数の正方形の画像を入力する必要があるため，グレースケールの値の割り当てが終了した箇所から 64 バイト目まで白画素（グレースケールの値：255）を付与する．この処理を 64 パケットに対して行い，64 × 64 画素の画像 1 枚に変換する．トラフィックデータを画像に変換した例を**図 4** に示す．図中の 1 行が 1 パケットを表しており，画像全体で 64 パケットが表されている．右端が白画素で埋められていることも確認できる．このようにトラフィックデータを画像に変換することでトラフィックの特徴を顕著に表すことができる．

3.2 VAE による学習

学習期間のトラフィックデータを 3.1 節の手法により画像に変換し，学習用データとして VAE で学習する．VAE は AE と比較して，潜在変数の表現に確率分布を用いていることが大きな特徴である．VAE の概要を**図 5** に示す．VAE は Encoder と Decoder の 2 つで構成され，Encoder で入力データの次元数を削減し，Decoder で削減された次元のデータから元の画像を再構成する．その際に入力データ $\boldsymbol{x}$ を潜在変数 $\boldsymbol{z}$ に確率分布 $p(\boldsymbol{z})$ として写像する．入力から潜在変数へ直接変換するのではなく，潜在変数はある確率分布からサンプリングされると仮定し，その平均と分散を Encoder で求める．しかし，このモデルをそのまま適用すると誤差逆伝播法で学習できないため，式（1）の近似式を用いる．式（1）では，$\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ という確率変数を導入し，μ は平均，σ は標準偏差を示している．

$$z = \boldsymbol{\mu} + \varepsilon \boldsymbol{\sigma}^2 \tag{1}$$

また，VAE は式（2）に示す損失関数を用いている．VAE の目的は周辺尤度 $p(\boldsymbol{x})$ の最大化であり，$p(\boldsymbol{x})$ が最大化されるとき，$\log p(\boldsymbol{x})$ も最大化される．ここでは，潜在変数の真の分布 $p(\boldsymbol{z} \mid \boldsymbol{x})$ を近似する分布を $q(\boldsymbol{z} \mid \boldsymbol{x})$ とする．また，$q(\boldsymbol{z} \mid \boldsymbol{x})$ は Decoder に $\boldsymbol{z}$ を入力したときにおける $\boldsymbol{x}$ の確率分布を表しており，$p(\boldsymbol{x} \mid \boldsymbol{z})$ は Encoder に $\boldsymbol{x}$ を入力したときにおける $\boldsymbol{z}$ の確率分布を表している．$D_{\mathrm{KL}}[q(\boldsymbol{z} \mid \boldsymbol{x})||p(\boldsymbol{z})]$ はカルバック・ライブラー情報量といい，事前分布 $p(\boldsymbol{z})$ と事後分布 $q(\boldsymbol{z} \mid \boldsymbol{x})$ の類似度を評価している．式（2）によって示される周辺尤度の下限 $L = -D_{\mathrm{KL}}[q(\boldsymbol{z} \mid \boldsymbol{x})||p(\boldsymbol{z})] + E_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x} \mid \boldsymbol{z})]$ は変分下限（ELBO：evidence lower bound）と呼ばれる．VAE の学習は，この変分下限を最大化するようにパラメータの更新が行われる．

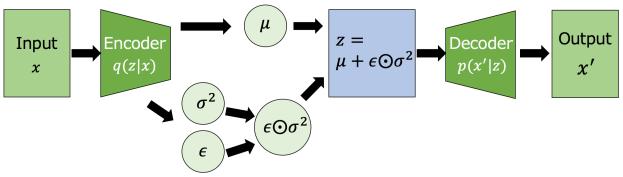

図 5 VAE の構成
Fig. 5 Constitution of Variational AutoEncoder.

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} \\
&= \log \int q(\boldsymbol{z} \mid \boldsymbol{x}) \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x})} d\boldsymbol{z} \\
&\geq \int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x})} d\boldsymbol{z} \\
&= \int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x})} d\boldsymbol{z} \\
&= \int q(\boldsymbol{z} \mid \boldsymbol{x}) \left(-\log \frac{q(\boldsymbol{z} \mid \boldsymbol{x})}{p(\boldsymbol{z})} + \log p(\boldsymbol{x} \mid \boldsymbol{z}) \right) d\boldsymbol{z} \\
&= -D_{\mathrm{KL}}[q(\boldsymbol{z} \mid \boldsymbol{x})||p(\boldsymbol{z})] + E_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x} \mid \boldsymbol{z})]
\end{aligned} \tag{2}$$

3.3 異常検出

3.2 節で学習した VAE を用いて異常を検出する．正常なデータのみを用いて VAE で学習すると，学習後の VAE は正常なデータの確率分布に従って画像を再構成する．したがって，学習後の VAE にテスト用画像 $\boldsymbol{x}$ を入力した際に，$\boldsymbol{x}$ が正常なデータの画像であれば，正常なデータの確率分布に従って $\boldsymbol{x}$ に似た画像 $\boldsymbol{x'}$ を再構成することができる．一方，正常なデータの確率分布に従わない $\boldsymbol{x}$ が入力された場合には，$\boldsymbol{x}$ に似た画像 $\boldsymbol{x'}$ を再構成することができず，学習用データに含まれない異常なデータの画像であると判断できる．以上から，式（3）を用いて入力した画像 $\boldsymbol{x}$ と再構成した画像 $\boldsymbol{x'}$ の差分を異常度 $A(\boldsymbol{x})$ として算出する．$A(\boldsymbol{x})$ が閾値未満ならば正常，閾値以上ならば異常とし，正常な通信と異常な通信を識別する．

$$A(\boldsymbol{x}) = \sum \left| \boldsymbol{x} - \boldsymbol{x'} \right| \tag{3}$$

4. 実験

本手法の有効性を確認する実験を行った．評価方法は ROC（Receiver Operating Characteristic）曲線及び AUC（Area Under the Curve）値を用いた．ROC 曲線とは正常と異常を分類する際の閾値を変化させたときの真陽性（True positive Rate）と偽陽性（False Positive Rate）を計算し，プロットしたときの軌跡である．ここで，真陽性とは異常データに対して正しく異常とした割合を示し，偽陽性とは正常データを誤って異常と検出した割合を示す．すなわち ROC 曲線が左上方に近づくほど検出性能が高いことを示している．また ROC 曲線が左上方に近づいている

表 1　BOS データセットの画像数
Table 1　Number of images in BOS dataset.

データセット	日付	進行度	正常	異常	合計
学習用	2017/8/17	2	10862	0	10862
学習用	2017/8/18	2	9416	0	9416
テスト用	2018/1/23	8	140	55	195
テスト用	2018/1/24	8	2549	216	2765

ことを表す指標として AUC 値を用いる．AUC 値は ROC 曲線の下側の面積を算出したものであり，この値が 1 に近づくほど完全に分類できていることを示す．

4.1　実験条件

実験には MWS 2018 の BOS データセット [10] と CICIDS 2017 データセット [11] を使用した．データセットの詳細については次節以降で説明する．テスト用データに対する正常と異常のラベル付けでは，C2 サーバ（Command & Contral サーバ）または攻撃者端末と通信しているパケットを変換した画像に異常のラベル，それ以外に正常のラベルを付与することとした．また，バッチサイズは 64 とし，学習回数は学習誤差の遷移を確認し，データセット毎に決定した．

4.2　実験データセット

4.2.1　BOS データセット

MWS 2018 の BOS データセット [10] は，標的型攻撃のマルウェアをローカルネットワーク内の端末に感染させ，その通信をキャプチャした研究用データセットである．BOS データセットはマルウェアとの通信の進行度によって 1 から 8 まで定義されている．進行度 1，2 のデータはマルウェアを実行したが C2 サーバとの通信は発生しておらず，進行度 3，4，5 のデータは C2 サーバとの通信は発生したが C2 サーバとの通信は成立していない状態を示している．進行度 6，7，8 のデータは通信が発生し，C2 サーバとの通信も成立している状態である．実験では，2017/8/17 ～ 2017/8/18 に収集された進行度 2 のトラフィックデータを異常が含まれていない正常なトラフィックデータとして学習に用いた．また，2018/1/23 ～ 2018/1/24 に収集された進行度 8 のトラフィックデータを異常を含むトラフィックデータとしてテストに用いた．本実験で用いた学習用データの画像は 20278 枚，テスト用データの画像は 2960 枚である．表 1 に，ラベル付けした学習用データとテスト用データの画像枚数の詳細を示す．

4.2.2　CICIDS データセット

CICIDS 2017 データセット [11] は，攻撃者端末から組織内ネットワークへのサイバー攻撃を想定した研究用データセットである．2017/7/3 ～ 2017/7/7 までのトラフィックデータをキャプチャしており，7/3 のトラフィックは全て正常な通信である．それ以外の日の通信は攻撃通信と正常通信の両方を含むトラフィックデータである．学習用データとして 7/3 のトラフィックデータを使用した．テスト用データとしては，それ以外の日のトラフィックデータを使用した．テスト用データの中には正常通信に加えて，攻撃通信として 7 種類の攻撃（DoS / DDoS，ポートスキャン，ブルートフォースアタック，クロスサイトスクリプティング，Botnet，cooldisk，Dropbox）が発生した時の通信が含まれている．表 2 に各手法毎 に正常および異常のラベル付けをした学習用データとテスト用データの画像枚数の詳細をそれぞれ示す．

表 2　CICIDS データセットの画像数
Table 2　Number of images in CICIDS dataset.

データセット	検知対象	正常	異常	合計
学習用	-	100000	-	100000
テスト用	DoS/DDoS	29999	19776	49775
	ブルートフォース	13179	4393	17572
	クロスサイト スクリプティング	438	146	584
	Portscan	15120	5040	20160
	Botnet	510	170	680
	cooldisk	687	229	916
	Dropbox	2049	683	2732

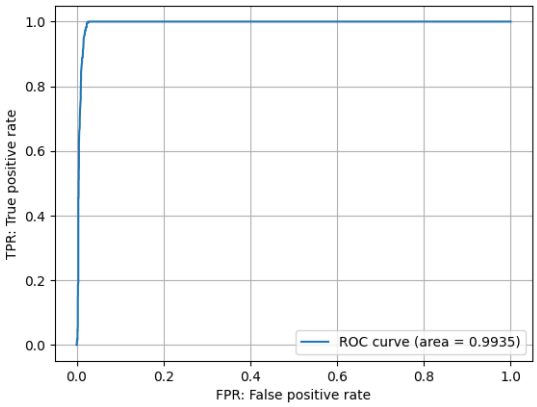

図 6　BOS データセットを用いたときの ROC 曲線
Fig. 6　ROC curve of Experiment using BOS dataset.

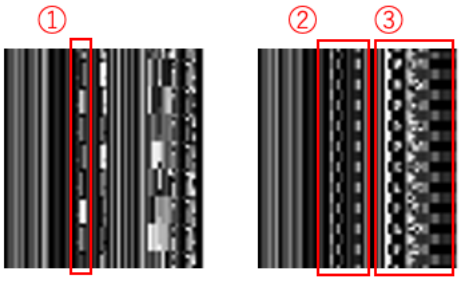

図 7　学習用データの例（BOS データセット）
Fig. 7　Examples of training dataset(BOS dataset).

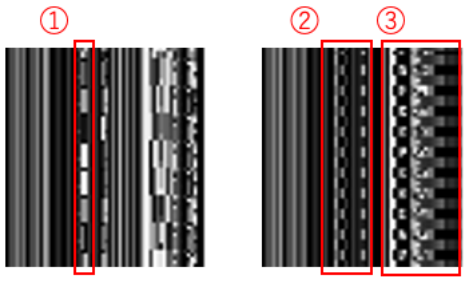

図 8　テスト用データ中の正常画像の例（BOS データセット）
Fig. 8　Examples of normal images in test dataset(BOS dataset).

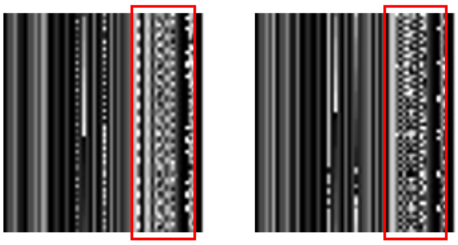

図 9　テスト用データ中の異常画像の例（BOS データセット）
Fig. 9　Examples of abnormal images in test dataset(BOS dataset).

4.3　実験結果と考察

4.3.1　BOS データセットを用いた実験結果

BOS データセットを用いて学習した際のエポック数に対する学習誤差の遷移を確認し，60 エポック程度で誤差の遷移が緩やかになり徐々に収束していることがわかった．そこで，誤差の遷移の収束が確認できるエポック数 60 を本実験で使用するエポック数として設定した．図 6 に本手法の ROC 曲線および AUC 値を示す．本手法の AUC 値は 0.9935 と高い値が得られ，本手法の有効性を確認できた．図 7 に学習用データの画像の例を示し，図 8，図 9 にテスト用データ中の正常画像の例と異常画像の例を示す．図 7 の画像と図 8 の画像の赤枠で囲った①～③の箇所をそれぞれ比較すると，類似する縦線が等間隔に現れている．一方で，図 9 の画像には赤枠で囲った箇所に等間隔に縦線が現れているものの図 7 のような縦線の現れ方とは異なる．VAE が図 7 のような学習用データの特徴を学習したため，図 8 のような正常画像が入力された際には，類似する画像を再構成できたが，図 9 のような異常画像が入力された際には類似する画像を再構成できなかったと考えられる．その結果として，AUC 値が高くなったと考えられる．

正常なテスト用データの異常度と異常なテスト用データの異常度の分布を確認するために，異常度のヒストグラムを作成した．作成したヒストグラムを図 10 に示す．図 10 では，ヒストグラムの縦軸を出現頻度，横軸を異常度の値としている．ほとんどの正常なトラフィックデータから算出した異常度は低くなり，異常なトラフィックデータから

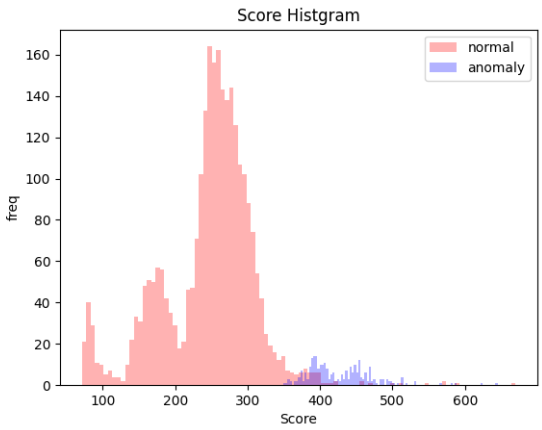

図 10　異常度のヒストグラム（BOS データセット）
Fig. 10　Histogram of Anomaly Scores(BOS dataset).

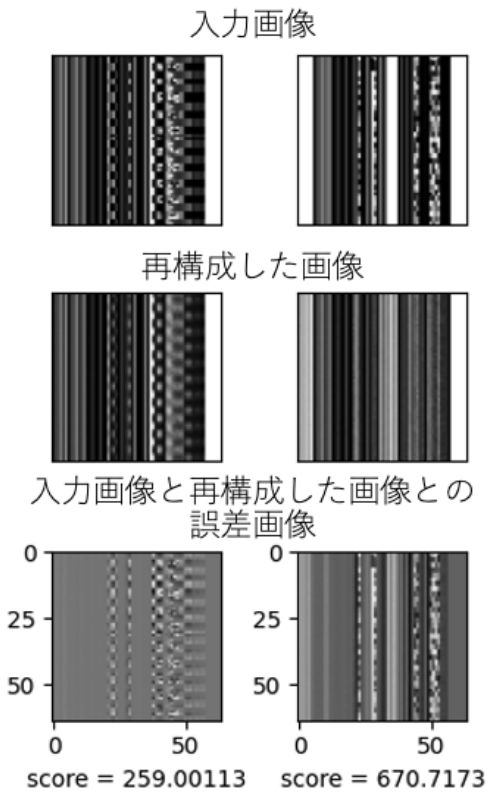

図 11　誤差画像の比較（左：異常度が低かったテスト用データの正常画像　右：異常度が高かったテスト用データの正常画像）
Fig. 11　Comparison of Difference images (Left : Normal image of test data with low Anomaly Score Right : Normal image of test data with high Anomaly Score)．

算出した異常度は高くなっていることがわかる．しかし，正常なトラフィックデータから算出した異常度でも高い値を示す場合が存在した．そこで，異常度が低かった場合と異常度が高かった場合の正常画像を比較した．図 11 に VAE に入力した画像と入力された画像から再構成した画像及びその誤差画像を示す．左の列の画像が異常度が低かった例，右の列の画像が異常度が高かった例を示している．

表 3　CICIDS データセットを用いたときの実験結果
Table 3　Experimental Results using CICIDS dataset.

検知対象	AUC 値
DoS/DDoS	0.9994
ブルートフォース	0.9959
クロスサイト スクリプティング	0.9997
Portscan	0.9997
Botnet	0.8710
cooldisk	0.7789
Dropbox	0.7364

異常度が高くなった要因として，図 11 の右上の画像は学習用データの画像と類似しておらず，入力された画像に対する潜在変数を VAE が正確に求められなかったことが原因であると考えられる．これは，学習用データを増加させ，VAE が再構成できる画像の種類を増やすことで改善できると考えられる．

4.3.2　CICIDS データセットを用いた実験結果

前節と同様に学習誤差の遷移を確認し，75 エポック程度で誤差の遷移が緩やかになり徐々に収束していることがわかった．そこで，誤差の遷移の収束が確認できるエポック数 75 を本実験で使用するエポック数として設定した．表 2 に各攻撃毎の AUC 値を示す．cooldisk，Dropbox 以外の攻撃で AUC 値が 0.8 以上となっており，本手法の有効性を確認できた．

ブルートフォース攻撃の異常画像やボットネット攻撃の異常画像をテスト用データの正常画像と比較した画像を図 12，図 13 に示す．異常画像を確認すると，赤枠で囲った箇所で同様のパケットが等間隔に現れている．正常画像で赤線で囲った箇所と同様の箇所を確認すると等間隔でパケットが現れているような特徴は存在せず，異常画像とは明確に異なっている．学習用データには赤枠で囲った箇所のような特徴は存在しなかったため，VAE がその特徴を再構成できず差分が大きくなったと考えられる．

cooldisk の実験結果は AUC 値 0.7789 となり他の攻撃より低い値となった．低くなった要因としては，異常画像であるにもかかわらず，異常度が低く算出された画像の枚数が他の攻撃よりも多かったことが挙げられる．そこで，正常画像と異常度が低い cooldisk の画像，異常度が高い cooldisk の画像を比較した．テスト用データの正常画像と cooldisk の画像との比較画像を図 14 に示す．上段は正常画像，中段は異常度が低い cooldisk の画像，下段は異常度が高い cooldisk の画像である．上段と中段の画像を比較すると，赤枠で囲った箇所①の特徴が正常画像に現れる特徴と類似していることがわかる．その為に，VAE が類似する画像を再構成でき，異常度が低くなったと考えられる．上段と下段の画像を比較すると，赤線で囲われた箇所②③が正常画像に現れる特徴と大きく異なっている．学習用データ

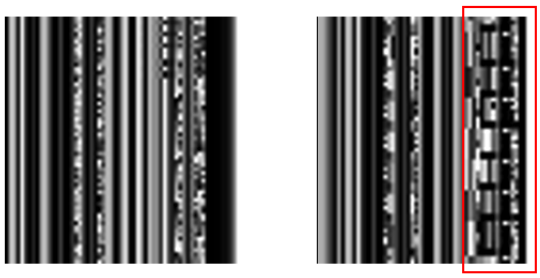

図 12　入力画像の比較（左：学習用データ　右：ブルートフォース攻撃発生時の画像）
Fig. 12　Comparison of Input images（Left : Normal image Right : Image of Brute Force attack）.

に赤線で囲われた箇所の特徴②③のような特徴が表れている画像は多く存在しておらず，VAE が再構成できなかったため，異常度が高くなったと考えられる．cooldisk は USB から感染したファイルが Macintosh マシンにダウンロードされた後，ポートスキャンの一種である Nmap（Network Mapper）をネットワーク全体で実行する攻撃である．異常度が低く算出された cooldisk の画像は，Nmap が起動しておらず，正常な通信のみが画像に変換され，正常画像と類似していた可能性がある．今後，異常通信の画像の精査を行いたいと考えている．

Dropbox の実験結果も AUC 値 0.7364 となり他の攻撃より低い値となった．低くなった要因として cooldisk の時と同様に，異常画像であるにもかかわらず，異常度が低く算出された画像の枚数が他の攻撃よりも多かったことが挙げられる．Dropbox はマルウェアに感染したファイルが Dropbox から Windows マシンにダウンロードされた後，ポートスキャンの一種である Nmap をネットワーク全体で実行する攻撃である．異常度が低く算出された cooldisk の時と同様に，Nmap が起動しておらず，正常な通信のみが画像に変換され，正常画像と類似していた可能性がある．こちらについても異常通信の画像の精査を行いたいと考えている．

4.3.1 節での BOS データセットを用いた実験結果と本節での実験結果から，本手法の有効性を確認できた．また，宛先 IP アドレスと送信元 IP アドレスの組み合わせ毎に観測されたパケットを画像に変換する手法を用いることで，画像に様々なサイバー攻撃の特徴が顕著に現れることも確認できた．一方，正常画像であるにもかかわらず異常度が高くなる画像や cooldisk や Dropbox のように異常画像であるにもかかわらず異常度が低くなる画像も存在した．今後の課題としては，正常な画像の種類の増加手法の検討や異常通信の精査等が挙げられる．

5.　まとめ

本論文では，深層学習手法の一つである VAE を用いて，

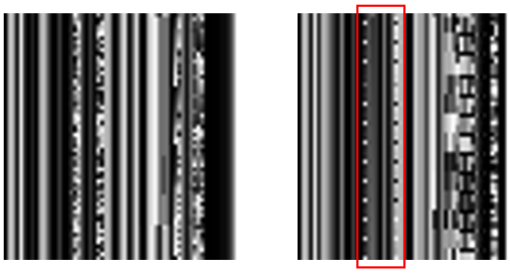

図 13　入力画像の比較（左：学習用データ　右：Botnet 攻撃発生時の画像）

Fig. 13　Comparison of Input images（Left : Normal image Right : Image of Botnet attack）.

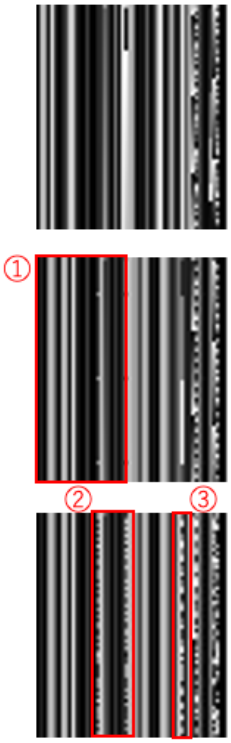

図 14　入力画像の比較（上段：学習用データ　中段：異常度が低かった cooldisk 発生時の画像　下段：異常度が高かった cooldisk 発生時の画像）

Fig. 14　Comparison of Input images（Top : Normal image Middle : Image of cooldisk with low Anomaly Score Bottom : image of cooldisk with high Anomaly Score）.

ネットワークトラフィックの異常を検出する手法を提案した．実験では，2 種類のデータセットを用いて本手法の有効性を確認した．今後の課題としては，正常な画像の種類の増加手法の検討や異常通信と画像の精査等が挙げられる．

参考文献

[1]　Snort, <https://www.snort.org/>（参照 2021-08-12）.

[2]　Suricata, <http://suricata-ids.org/>（参照 2021-08-12）.

[3]　The Bro, <http://www.bro.org/>（参照 2021-08-12）.

[4]　小島俊輔，中嶋卓雄，末吉敏則：エントロピーベースのマハラノビス距離による高速な異常検知手法，情報処理学会論文誌，Vol.52,，No.2，pp.656-668（2011）.

[5]　佐藤陽平，和泉勇治，根元義章：複数の検出モジュールの組み合わせによるネットワーク異常検出の高精度化，信学技報，NS2004-104，pp.45-48（2014）.

[6]　平松尚利，和泉勇治，角田裕，根元義章：複数の通常状態を用いたネットワーク異常検出，信学技報，CS2006-32，pp.61-66（2006）.

[7]　Hinton，Geoffrey E.，Ruslan R.，Salakhutdinov. ：Reducing the dimensionality of data with neural networks. Vol.313，pp.504-507（2006）.

[8]　Kingma，Diederik P.，Max Welling. : Auto-encoding variational bayes. The International Conference on Learning Representations（2014）.

[9]　森雅也，中平勝子，高橋弘毅，田中貴浩：ノイズに埋もれた微小信号波形の検出への変分オートエンコーダを用いた異常検知の応用，IEICE Conferences Archives，The Institute of Electronics（2019）.

[10]　高田雄太，寺田真敏，松木隆宏，笠間貴弘，荒木粧子，畑 田充弘：マルウェア対策のための研究用データセット MWS Datasets 2018，情報処理学会論文誌，Vol.2018- CSEC-82，No.38，pp.1-8（2018）.

[11]　I. Sharafaldin，A. Habibi Lashkari，A. A. Ghorbani : Toward generating a new intrusion detection dataset and intrusion traffic characterization.，Proc. 4th ICISSP，pp.108-116（2018）.