Vol.2020-CSEC-91 No.24
Vol.2020-SPT-40 No.24
Vol.2020-EIP-90 No.24
2020/11/26

ネットワークの分散表現学習に基づく亜種マルウェアの活動検知

岸波敬介[1]　梅澤猛[2]　大澤範高[2]

概要：マルウェア検知にあたっては，既存のマルウェアを改変した亜種マルウェアの検知が課題となる．本研究では，分散表現を利用することで，亜種検知率の高い推定モデル構築を目指した．通信トラフィックデータを対象に，パケットのヘッダ情報をフィールド単位に分割し，分散表現ベクトルを生成する．分散表現ベクトルの系列を Long short-term memory (LSTM)などによって解析することで，亜種マルウェアの活動を検知する．マルウェアに感染した機器のトラフィックデータとして Malrec Dataset，正常な機器の通信トラフィックデータとして Malware Capture Facility Project が提供している Normal Datasets と，NCD in MWS Cup 2014 を基にした実験において，訓練データに含まれない亜種マルウェアを正解率 0.97 で検知した．

キーワード：ネットワークトラフィック，マルウェア，亜種検出，LSTM，Self-Attention

1. はじめに

通信トラフィックの解析によるマルウェア検知は，IoT 機器など計算資源に乏しく，システムコール情報の取得やウイルス対策ソフトの導入・更新が困難な対象のセキュリティ対策に有効であるが，膨大なトラフィックの中から悪性なものを判別する検知モデル作成の自動化と効率化が課題である．

マルウェア検知にあたっては，既存のマルウェアを改変した亜種マルウェアを高い精度で検出できることが求められる．攻撃者にとって亜種の作成は全く新しいマルウェアを作成するよりも容易であるが，亜種のセキュリティソフトで広く使用されているパターンマッチング方式では，亜種の検出精度は充分とは言えない．そこで本研究では，分散表現を利用することで，亜種検知率の高いモデル構築を目指す．分散表現は単語や文書などを多次元ベクトルで表現する手法であり，自然言語処理分野でその有効性を高く評価されている．通信トラフィックを分散表現することで，亜種マルウェアのトラフィックが類似性の高い分散表現ベクトルとして表現され、機械学習によってマルウェアの活動に伴って発生する通信を判別するモデルを構築し，それを利用することで従来は検出が難しかった学習データに含まれない亜種マルウェアの検出が可能になることが期待される．

2. 関連研究

水野ら[1]は，マルウェアの活動において使用されることの多い HTTP 通信に着目し，そのヘッダ情報から特徴を抽出して悪性通信と良性通信の分類を行った．実データを利用した評価実験では，正解率 0.905 で分類が可能であることを示した．亜種の検出性能は評価されていない．

武部[2]は，マルウェアの動的解析結果から API コール列を抽出し，その API 関数名を実行された順番に並べて Paragraph Vector (PV)を作成し，Support Vector Machine (SVM)モデルでマルウェアの亜種推定を行った．その結果，平均正解率 0.840，平均 F 値 0.842 の性能で亜種推定を行えることを示した．本稿とは異なり，通信トラフィックではなく，感染端末で実行された API 関数の情報から特徴抽出を試みている．

3. 分散表現

分散表現は，自然言語処理分野で有効性が示された，単語や文章の密な実数ベクトル表現である[3].

電子計算機上で文章を扱うためには，単語の 1-of-K 符号化と文章の Bag-of-Words 表現が広く用いられてきた．1～K 番目までの単語を，K 列目の要素を 1 に，それ以外の要素を 0 にした K 次元のベクトルで表現するのが 1-of-K 符号化である．文章中に含まれる単語の 1-of-K ベクトルを足し合わせて，得られたベクトルをその文章の表現として扱う手法が Bag-of-Words である．しかし，Bag-of-Words による文章表現には，文章に含まれる語彙の増加とともに次元数もまた膨大なものとなる問題点，また文章中に含まれる単語の語順情報が失われる問題点が存在する．

そこで，1-of-K 符号化によって表現された高次元のベクトルを低次元のベクトルに変換して，語順情報を加味しながら単語間の意味の近さをも同時に表現する手法として分散表現が考案された．Erhan ら[4]は，機械学習において 1-of-K ベクトルのような高次元ベクトルを低次元の実数ベクトルに変換することで，より高い性能を発揮できると述べている．

単語を分散表現するための手法としては，Mikolov ら[5][6]の Continuous Bag-of-Words (CBOW) モデルや Skip-gram モデルを利用する word2vec や，Transformer モデルを利用する BERT [7]などがある．

文章を分散表現するための手法としては，Le ら[8]の Distributed Memory Model of Paragraph Vector (PV-DM) モデ

1 千葉大学 大学院融合理工学府
Graduate School of Science and Engineering, Chiba University.

2 千葉大学 大学院工学研究院
Graduate School of Engineering, Chiba University.

Vol.2020-CSEC-91 No.24
Vol.2020-SPT-40 No.24
Vol.2020-EIP-90 No.24
2020/11/26

ルや Distributed Bag of Words version of Paragraph Vector (PV-DBOW) モデルを利用する doc2vec や，BERT などがある．

4. 提案手法

通信トラフィックを分散表現学習に基づいてベクトル化し，機械学習によってマルウェア検知モデルを構築する．本研究では，パケットのヘッダ情報をモデル構築に利用する．現代の情報通信においてペイロードは暗号化されていることが多く，適切な情報抽出が困難であるためである．

パケットヘッダのフィールド名と，そのデータの組を「単語」と定義し，その分散表現を学習する．分散表現である単語ベクトルの列として表現された通信トラフィックを Long short-term memory (LSTM)などで時系列解析することで，マルウェア通信の特徴を学習した検知モデルを構築し，良性通信・悪性通信の分類とファミリ分類を行う．

5. データセット

悪性通信データセットとして，The Malrec Dataset [9]の Packet Capture (PCAP)ファイルを利用した．このデータセットは，2014/12/07 - 2016/12/03 の期間に Georgia Tech Information Security Center やプロバイダ，アンチウイルスベンダーから提供された総数 66,301 件のマルウェア検体に対して，サンドボックス解析システム Malrec [10]を用いて挙動を解析・記録したデータ群である．

The Malrec Dataset は自動分類手法 AVCLASS [11]に基づいてマルウェア名がラベリングされているが，その過程でマルウェア検体の亜種名の情報が失われている．そこで，Kaspersky [12]の分類名に基づき再度マルウェア検体のラベリングを行った．

Kaspersky は，マルウェアの命名規則を"[Prefix:]Behaviour.Platform.Name[.Variant]"と定めている[13]. Prefix は検体が検知されたサブシステム名，Behaviour はマルウェアの種類，Platform は検体の実行環境，Name はマルウェア名，Variant は亜種名を表す．

マルウェア検体の MD5 ハッシュ値をオンラインマルウェア解析サービス VirusTotal[14]へ送信し，Kaspersky が定めた分類名を取得する．本研究においては，"Behaviour.Platform.Name"を一つのマルウェアファミリとして扱い，"Variant"を亜種として扱う．

PCAP ファイルをネットワークプロトコルアナライザ tshark[15]で解析し，提案手法に基づいて単語に変換したテキストデータを，機械学習モデルの入力に用いた．

良性通信のデータセットとしては，MWS Datasets [16]に含まれる NCD in MWS Cup 2014 の 5 つの PCAP ファイル，ならびに Stratosphere Lab が提供する Normal Datasets [17]における CTU-normal-18，CTU-normal-20， CTU-normal-33 の PCAP ファイルを利用した．

表 1 に，提案手法に基づいて定義した単語の数と語彙数を示した．

表 1 データセットの性質

データセット名	語彙数	総単語数
The Malrec Dataset	3,287,668	119,719,067
NCD in MWS Cup 2014	823,979	9,356,485
Normal Datasets	439,269	3,748,432

6. 評価指標

本研究における評価指標としては，Accuracy (正解率)，Precision (適合率)，Recall (再現率)，F-measure (F1，F 値)の 4 つを用いる．

式(1)から式(7)に，表 2 の混同行列を基にした各指標の定義式を示す．iは評価するクラス，N は評価するクラス数である．多クラス分類の場合，評価するクラスごとに適合率，再現率，F 値を計算し，それらを平均した Macro-Precision (マクロ平均適合率)，Macro-Recall (マクロ平均再現率)，Macro-F1 (マクロ平均 F 値)を求める．

表 2 混同行列

		推定		
		1	… j …	N
正解	1	E_{11}	E_{1j}	$E_{1\mathrm{N}}$
	⋮ i ⋮	E_{i1}	⋮ … E_{ij} … ⋮	$E_{i\mathrm{N}}$
	N	$E_{\mathrm{N}1}$	… $E_{\mathrm{N}j}$ …	E_{NN}

$$\mathrm{Accuracy} = \frac{\sum_{i=1}^{\mathrm{N}} E_{ii}}{\sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{N}} E_{ij}} \quad \text{式(1)}$$

$$\mathrm{Precision}_{(i)} = \frac{E_{ii}}{\sum_{j=1}^{\mathrm{N}} E_{ji}} \quad \text{式(2)}$$

$$\mathrm{Recall}_{(i)} = \frac{E_{ii}}{\sum_{j=1}^{\mathrm{N}} E_{ij}} \quad \text{式(3)}$$

$$\mathrm{F1}_{(i)} = \frac{2 \times \mathrm{Precision}_{(i)} \times \mathrm{Recall}_{(i)}}{\mathrm{Precision}_{(i)} + \mathrm{Recall}_{(i)}} \quad \text{式(4)}$$

$$\text{Macro-Precision} = \frac{\sum_{i=1}^{\mathrm{N}} \mathrm{Precision}_{(i)}}{\mathrm{N}} \quad \text{式(5)}$$

$$\text{Macro-Recall} = \frac{\sum_{i=1}^{\mathrm{N}} \mathrm{Recall}_{(i)}}{\mathrm{N}} \quad \text{式(6)}$$

$$\text{Macro-F1} = \frac{\sum_{i=1}^{\mathrm{N}} \mathrm{F1}_{(i)}}{\mathrm{N}} \quad \text{式(7)}$$

7. 実験 1 – LSTM による分類

7.1 実験 1.a – 良性通信・悪性通信の分類

提案手法に基づいてマルウェア検知モデルを構築し，良性通信と悪性通信の分類精度を評価する．

表 3 に，良性通信と悪性通信の分類実験に使用した学習データの内訳を示した．良性通信のデータセットについて

Vol.2020-CSEC-91 No.24
Vol.2020-SPT-40 No.24
Vol.2020-EIP-90 No.24
2020/11/26

は，PCAP ファイルを 500 パケットごとに分割し，計 320 件を抽出した. 以降の実験でもすべて同様の処理を行った．悪性通信のデータについて，パケット列の長さは可変長であり，最大で 6441 パケット，最小で 38 パケット，平均は 440 パケット，標準偏差は 327.7 であった．悪性通信には (a)P2P-Worm.Win32.Sytro ， (b)Trojan.Win32.Reconyc ， (c)Virus.Win32.Expiro ， (d)Virus.Wiin32.PolyRansom ， (e)Worm.Win32.WBNA の 5 種類のファミリが含まれ，それぞれ 3 種類の亜種が存在する．1 種類の亜種を選択して未知亜種検証データとし，それ以外の 2 種類を訓練データと既知亜種検証データとした．訓練データで学習したモデルに対して検証用のデータを入力し，既知亜種と未知亜種の検出性能を評価する．

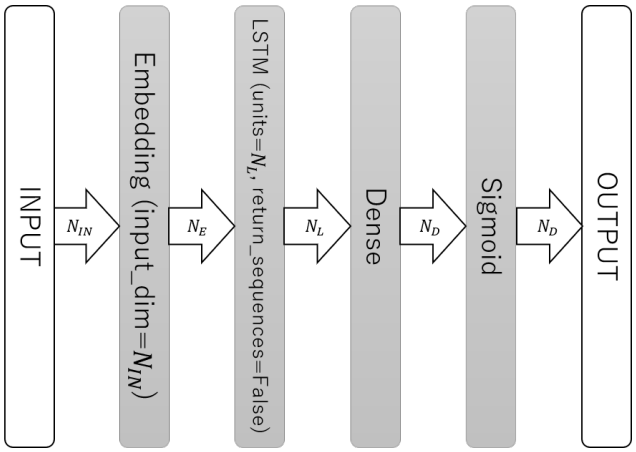

図 1　二値分類モデル (LSTM)

図 1 にモデルの構造を示す．矢印の内部に出力の次元数を示し，以降のモデルの構造図においても同様に表す．モデルは Python のニューラルネットワークライブラリ keras[18]を用いて構築した．このモデルは，通信トラフィックを変換して得た単語の系列データ$X(x_1, x_2, \ldots x_t)$を入力として受け取り, Embedding 層で単語$x_i(1 \le i \le t)$がN_E次元の単語ベクトル$\boldsymbol{v_i}(1 \le i \le t)$に変換され, LSTM 層に順に入力される．$x_i$は整数化(トークン化)されており，One-hot ベクトルとしての次元数は総語彙数$N_{IN} = 3{,}287{,}668$である．また，以降の全ての実験で$N_E = 50$とした．LSTM 層の内部ユニット数$N_L = 50$として，出力を Dense 層に入力する．Dense 層は出力ユニット数$N_D = 1$，活性化関数をシグモイド関数として, 良性通信と悪性通信の二値に分類する．

15 epoch の学習を行った結果を，表 4 に示す．(a).k，(b).ftbc，(c).ao，(d).b，(e).ipa のデータを未知亜種検証データとした．表では良性通信を「良」，悪性通信を「悪」と示した. 既知亜種を F 値 0.9587 で, 未知亜種を F 値 0.9691 で検知した．

7.2 実験 1.b – ファミリ分類

提案手法に基づいてファミリ分類モデルを構築し，良性通信と 5 クラスのマルウェアファミリについて分類精度を評価する．

表 5 に，ファミリ分類の実験に使用した学習データの内訳を示した．悪性通信には，実験 1 と同じ 5 種類のマルウェアファミリのデータが含まれる．ここからそれぞれ 1 種類の亜種を選択し, 40 件ずつを未知亜種の検証データとした．それ以外の亜種から 160 件ずつを選択し，訓練データと既知亜種の検証データとした．ただし，WBNA についてはデータセット内のサンプルが 160 件に満たないため, 130 件とした．訓練データで学習したモデルに対して検証用のデータを入力し，既知亜種と未知亜種の検出性能を評価する．

図 2 にモデルの構造を示す．単語の系列データXを入力として受け取り，良性通信と 5 種類のマルウェアファミリの計 6 クラスに分類する．Embedding 層，LSTM 層は 7.1 項と同じである$(N_E = 50, N_L = 50)$．Dense 層の中間層をユニット数$N_{D1} = 128$，出力層をユニット数$N_{D2} = 6$，活性化関数をソフトマックス関数として，良性通信と 5 種類のマルウェアファミリに分類する．

50 epoch の学習を行った結果を，表 6 から表 9 に示した．表 6，表 7，表 8 をみると，いずれも(b)のクラスに誤分類される傾向が強く見られた．この(b)クラスと良性通信を除いた 4 クラス分類を行い，表 10 から表 13 に結果を示した．訓練・既知亜種検証・未知亜種検証データのいずれも，(c)のクラスに誤分類される傾向が強く見られた．

表 3　良性・悪性の分類のためのデータの内訳 (LSTM)

データ	良性	悪性	悪性亜種内訳														
			(a)			(b)			(c)			(d)			(e)		
			j	k	o	cdcf	ftbc	gunk	ai	ao	ar	a	b	c	bul	ipa	ipi
訓練	157	154	18	13		18	13		18	13		18	13		17	13	
既知亜種検証	63	71	7	7		7	7		7	7		7	7		8	7	
未知亜種検証	100	100			20			20			20			20			20

Vol.2020-CSEC-91 No.24
Vol.2020-SPT-40 No.24
Vol.2020-EIP-90 No.24
2020/11/26

表 4　良性通信・悪性通信の分類結果 (LSTM)

訓練				既知亜種検証				未知亜種検証			
		推定				推定				推定	
		良	悪			良	悪			良	悪
正解	良	157	0	正解	良	58	5	正解	良	94	6
	悪	0	154		悪	0	71		悪	0	100
正解率		1.0000		正解率		0.9607		正解率		0.9700	
適合率		1.0000		適合率		1.0000		適合率		1.0000	
再現率		1.0000		再現率		0.9206		再現率		0.9400	
F 値		1.0000		F 値		0.9587		F 値		0.9691	

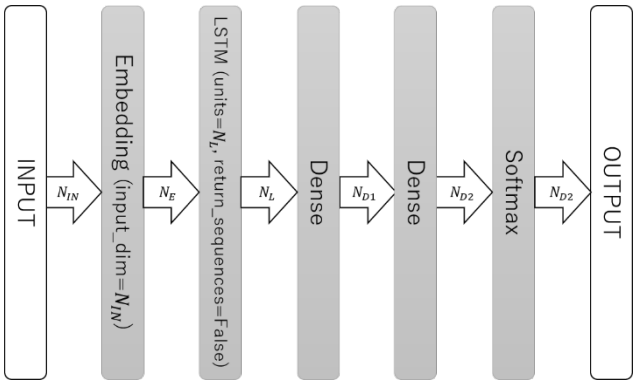

図 2　ファミリ分類モデル (LSTM)

表 5　ファミリ分類のためのデータの内訳

データ	良性	悪性	悪性亜種内訳																				
			(a)			(b)				(c)						(d)					(e)			
			j	o	k	cdcf	ftbc	gunk	gste	ai	ao	ar	ns	nt	w	a	b	c	f	k	bul	ipa	roc	ipi
訓練	120	577	60	60		40	40	40		25	25	21	25	24		35	35	15	35		50	40	7	
既知亜種検証	40	193	40	40		13	13	14		8	8	8	8	8		10	10	10	10		20	10	3	
未知亜種検証	40	200			40				40						40					40				40

表 6　ファミリ分類結果 – 訓練 (LSTM)

		推定						Recall
		良	(a)	(b)	(c)	(d)	(e)	
正解	良	120	0	0	0	0	0	1.0000
	(a)	0	18	102	0	0	0	0.1500
	(b)	0	0	119	1	0	0	0.9917
	(c)	0	0	104	16	0	0	0.1333
	(d)	0	0	54	0	66	0	0.5500
	(e)	0	0	81	0	0	16	0.1649
Precision		1.0000	1.0000	0.2587	0.9412	1.0000	1.0000	

表 7　ファミリ分類結果 – 既知亜種検証 (LSTM)

		推定						Recall
		良	(a)	(b)	(c)	(d)	(e)	
正解	良	40	0	0	0	0	0	1.0000
	(a)	0	7	32	1	0	0	0.1750
	(b)	0	0	40	0	0	0	1.0000
	(c)	0	1	34	0	5	0	0.0000
	(d)	0	1	18	0	21	0	0.5250
	(e)	0	0	27	0	4	2	0.0606
Precision		1.0000	0.7778	0.2649	0.0000	0.7000	1.0000	

表 8　ファミリ分類結果 – 未知亜種検証 (LSTM)

		推定						Recall
		良	(a)	(b)	(c)	(d)	(e)	
正解	良	39	0	0	0	1	0	0.9750
	(a)	0	4	36	0	0	0	0.1000
	(b)	0	0	40	0	0	0	1.0000
	(c)	0	0	36	0	3	1	0.0000
	(d)	0	0	35	0	5	0	0.1250
	(e)	0	1	0	0	39	0	0.0000
Precision		1.0000	0.8000	0.2721	0.0000	0.1012	0.0000	

表 9　評価指標 (LSTM)

訓練		既知亜種検証		未知亜種検証	
Accuracy	0.5093	Accuracy	0.4721	Accuracy	0.3667
Macro-Precision	0.8666	Macro-Precision	0.6238	Macro-Precision	0.3627
Macro-Recall	0.4983	Macro-Recall	0.4601	Macro-Recall	0.3667
Macro-F1	0.6328	Macro-F1	0.5296	Macro-F1	0.3647

Vol.2020-CSEC-91 No.24
Vol.2020-SPT-40 No.24
Vol.2020-EIP-90 No.24
2020/11/26

表 10　ファミリ分類結果 – 訓練 (LSTM)

		推定				Recall
		(a)	(c)	(d)	(e)	
正解	(a)	18	102	0	0	0.1500
	(c)	0	120	0	0	1.0000
	(d)	0	54	65	1	0.5417
	(e)	0	81	0	16	0.1649
Precision		1.0000	0.3361	1.0000	0.9412	

表 11　ファミリ分類結果 – 既知亜種検証 (LSTM)

		推定				Recall
		(a)	(c)	(d)	(e)	
正解	(a)	6	33	1	0	0.1500
	(c)	1	34	3	2	0.8500
	(d)	1	18	21	0	0.5250
	(e)	0	27	2	4	0.1212
Precision		0.7500	0.3036	0.7778	0.6667	

表 12　ファミリ分類結果 – 未知亜種検証 (LSTM)

		推定				Recall
		(a)	(c)	(d)	(e)	
正解	(a)	4	36	0	0	0.1000
	(c)	0	36	3	1	0.9000
	(d)	0	35	5	0	0.1250
	(e)	0	0	40	0	0.0000
Precision		1.0000	0.3364	0.1042	0.0000	

表 13　評価指標 (LSTM)

訓練		既知亜種検証		未知亜種検証	
Accuracy	0.4792	Accuracy	0.4248	Accuracy	0.2813
Macro-Precision	0.8193	Macro-Precision	0.6245	Macro-Precision	0.3602
Macro-Recall	0.4642	Macro-Recall	0.4116	Macro-Recall	0.2813
Macro-F1	0.5930	Macro-F1	0.4961	Macro-F1	0.3158

8. 実験 2 – 双方向 LSTM モデルによる分類

8.1 実験 2.a – 良性通信・悪性通信の分類

双方向 LSTM (BiLSTM)モデルをベースとした推定モデルを設計し，検知性能の評価実験を行った．図 3 にモデルの構造を示す．このモデルは，通信トラフィックを変換して得た単語の系列データXを入力として受け取り，良性・悪性の二値分類を行う．Embedding 層は 7.1 項と同じである($N_E = 50$)．順方向 LSTM 層に$\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots \boldsymbol{v_t}$を時系列順に，逆方向 LSTM 層に$\boldsymbol{v_t}, \boldsymbol{v_{t-1}}, \ldots \boldsymbol{v_1}$を時系列と逆順に入力し，それぞれの出力を連結して Dense 層に入力する($N_L = 50$)．Dense 層の中間層をユニット数$N_{D1} = 128$，出力層をユニット数$N_{D2} = 1$，活性化関数をシグモイド関数として，良性通信と悪性通信の二値に分類する．

表 14 に学習に使用したデータの内訳を示す. 表 15 に，(a).o，(b).gunk，(c).ar，(d).c，(e).ipi を未知亜種検証データとして 10 epoch の学習を行った結果を示した．既知亜種を F 値 0.9915 で，未知亜種を F 値 1.0000 で検知した．

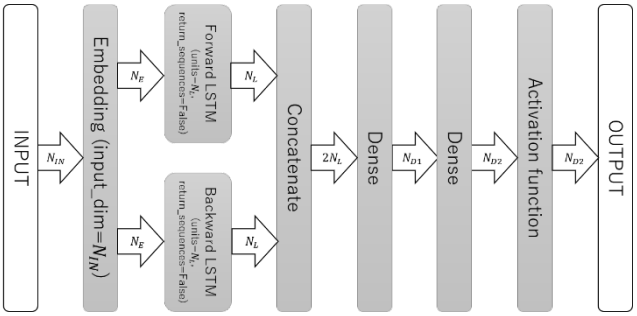

図 3　分類モデル (BiLSTM)

8.2 実験 2.b – ファミリ分類

双方向 LSTM (BiLSTM)モデルをベースとした推定モデルを設計し，良性通信と 5 クラスのマルウェアファミリについて分類精度を評価する．図 3 にモデルの構造を示す．このモデルは，単語の系列データXを入力として受け取り，良性通信と 5 種類のマルウェアファミリの計 6 クラスに分類する．モデルの構造は 8.1 項と同じであるが，$N_{D2} = 6$，活性化関数をソフトマックス関数として，良性通信と 5 種類のマルウェアファミリに分類する．

表 5 に，ファミリ分類の実験に使用した学習データの内訳を示した．50 epoch の学習を行った結果を，表 16 から表 19 に示した．表 9 と表 19 を比較すると訓練データの分類精度は高くなったが，検証データでは実験 1.b で見られたような，特定のクラス(c)に偏って分類される傾向が見られた．

Vol.2020-CSEC-91 No.24
Vol.2020-SPT-40 No.24
Vol.2020-EIP-90 No.24
2020/11/26

表 14　良性・悪性の分類のためのデータの内訳 (BiLSTM)

データ	良性	悪性	悪性亜種内訳													
			(a)			(b)			(c)			(d)			(e)		
			j	k	o	cdcf	ftbc	gunk	ai	ao	ar	a	b	c	bul	ipa	ipi
訓練	162	149	17	12		17	13		17	13		17	13		17	13	
既知亜種検証	58	76	8	8		8	7		8	7		8	7		8	7	
未知亜種検証	100	100			20			20			20			20			20

表 15　良性通信・悪性通信の分類結果 (BiLSTM)

訓練				既知亜種検証				未知亜種検証			
		推定				推定				推定	
		良	悪			良	悪			良	悪
正解	良	162	0	正解	良	58	0	正解	良	100	0
	悪	0	149		悪	1	75		悪	0	100
正解率		1.0000		正解率		0.9925		正解率		1.0000	
適合率		1.0000		適合率		0.9831		適合率		1.0000	
再現率		1.0000		再現率		1.0000		再現率		1.0000	
F 値		1.0000		F 値		0.9915		F 値		1.0000	

表 16　ファミリ分類結果 – 訓練 (BiLSTM)

		推定						Recall
		良	(a)	(b)	(c)	(d)	(e)	
正解	良	120	0	0	0	0	0	1.0000
	(a)	0	120	0	0	0	0	1.0000
	(b)	0	0	120	0	0	0	1.0000
	(c)	0	0	0	120	0	0	1.0000
	(d)	0	0	0	0	120	0	1.0000
	(e)	0	0	0	0	0	97	1.0000
Precision		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	

表 17　ファミリ分類結果 – 既知亜種検証 (BiLSTM)

		推定						Recall
		良	(a)	(b)	(c)	(d)	(e)	
正解	良	39	0	0	0	1	0	0.9750
	(a)	0	16	0	17	0	7	0.4000
	(b)	0	3	0	33	0	4	0.0000
	(c)	0	7	0	18	5	10	0.4500
	(d)	0	0	0	3	23	14	0.5750
	(e)	0	4	0	5	3	21	0.6364
Precision		1.0000	0.5333	0.0000	0.2368	0.7188	0.3750	

表 18　ファミリ分類結果 – 未知亜種検証 (BiLSTM)

		推定						Recall
		良	(a)	(b)	(c)	(d)	(e)	
正解	良	37	0	0	0	1	2	0.9250
	(a)	0	14	0	18	0	8	0.3500
	(b)	0	1	0	34	0	5	0.0000
	(c)	1	4	0	20	3	12	0.5000
	(d)	0	2	0	10	4	24	0.1000
	(e)	0	2	0	0	35	3	0.0750
Precision		0.9737	0.6087	0.0000	0.2439	0.0930	0.0556	

表 19　評価指標 (BiLSTM)

訓練		既知亜種検証		未知亜種検証	
Accuracy	1.0000	Accuracy	0.5021	Accuracy	0.3250
Macro-Precision	1.0000	Macro-Precision	0.4773	Macro-Precision	0.3291
Macro-Recall	1.0000	Macro-Recall	0.5061	Macro-Recall	0.3250
Macro-F1	1.0000	Macro-F1	0.4913	Macro-F1	0.3271

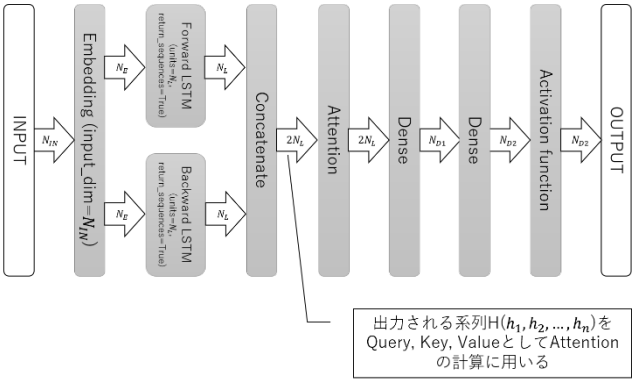

図 4　分類モデル (BiLSTM+Attention)

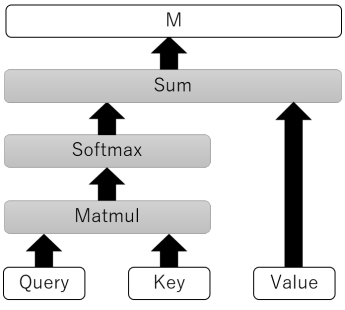

図 5　Attention 層

Vol.2020-CSEC-91 No.24
Vol.2020-SPT-40 No.24
Vol.2020-EIP-90 No.24
2020/11/26

表 20　良性・悪性の分類のためのデータの内訳 (BiLSTM+Attention)

データ	良性	悪性	悪性亜種内訳														
			(a)			(b)			(c)			(d)			(e)		
			j	k	o	cdcf	ftbc	gunk	ai	ao	ar	a	b	c	bul	ipa	ipi
訓練	149	162	18	14		18	14		18	14		19	14		19	14	
既知亜種検証	71	63	7	6		7	6		7	6		6	6		6	6	
未知亜種検証	100	100			20			20			20			20			20

表 21　良性通信・悪性通信の分類結果
(BiLSTM+Attention)

訓練				既知亜種検証				未知亜種検証			
		推定				推定				推定	
		良	悪			良	悪			良	悪
正解	良	149	0	正解	良	71	0	正解	良	100	0
	悪	0	162		悪	3	60		悪	1	99
正解率		1.0000		正解率		0.9776		正解率		0.9950	
適合率		1.0000		適合率		0.9594		適合率		0.9901	
再現率		1.0000		再現率		1.0000		再現率		1.0000	
F 値		1.0000		F 値		0.9793		F 値		0.9950	

表 22　ファミリ分類結果 – 訓練 (BiLSTM+Attention)

		推定						Recall
		良	(a)	(b)	(c)	(d)	(e)	
正解	良	120	0	0	0	0	0	1.0000
	(a)	0	120	0	0	0	0	1.0000
	(b)	0	0	118	0	0	2	0.9833
	(c)	0	112	0	6	0	2	0.0500
	(d)	2	0	0	0	118	0	0.9833
	(e)	0	6	1	2	0	88	0.9072
Precision		0.9836	0.5042	0.9916	0.7500	1.0000	0.9565	

表 23　ファミリ分類結果 – 既知亜種検証
(BiLSTM+Attention)

		推定						Recall
		良	(a)	(b)	(c)	(d)	(e)	
正解	良	38	0	0	0	2	0	0.9500
	(a)	0	29	2	4	0	5	0.7250
	(b)	0	4	24	1	0	11	0.6000
	(c)	0	12	8	4	1	15	0.1000
	(d)	0	2	12	1	23	2	0.5750
	(e)	0	11	10	2	1	9	0.2727
Precision		1.0000	0.5000	0.4286	0.3333	0.8519	0.2143	

表 24　ファミリ分類結果 – 未知亜種検証
(BiLSTM+Attention)

		推定						Recall
		良	(a)	(b)	(c)	(d)	(e)	
正解	良	35	0	0	0	5	0	0.8750
	(a)	0	29	1	4	0	6	0.7250
	(b)	0	0	33	0	0	7	0.8250
	(c)	0	12	8	2	4	14	0.0500
	(d)	4	2	9	0	3	22	0.0750
	(e)	0	10	22	1	0	7	0.1750
Precision		0.8974	0.5472	0.4521	0.2857	0.2500	0.1250	

表 25　評価指標 (BiLSTM+Attention)

訓練		既知亜種検証		未知亜種検証	
Accuracy	0.8178	Accuracy	0.5451	Accuracy	0.4542
Macro-Precision	0.8643	Macro-Precision	0.5547	Macro-Precision	0.4262
Macro-Recall	0.8206	Macro-Recall	0.5371	Macro-Recall	0.4542
Macro-F1	0.8419	Macro-F1	0.5458	Macro-F1	0.4398

Vol.2020-CSEC-91 No.24
Vol.2020-SPT-40 No.24
Vol.2020-EIP-90 No.24
2020/11/26

9. 実験 3 – Attention モデルによる分類

9.1 実験 3.a – 良性通信・悪性通信の分類

BiLSTM と Attention モデルをベースとした推定モデルを設計し，検知性能の評価実験を行った．図 4 にモデルの構造を示す．8.1 項のモデルに Attention 層を追加して，活性化関数をシグモイド関数とした．Attention 層は，Lin ら[19]の Self-Attention による文章の埋め込み手法を参考とした．図 5 に，Attention 層の構造を示す．順方向 LSTM 層の出力系列 $(\boldsymbol{h_1^f}, \boldsymbol{h_2^f}, \ldots \boldsymbol{h_t^f})$ と逆方向 LSTM 層の出力系列 $(\boldsymbol{h_1^b}, \boldsymbol{h_2^b}, \ldots \boldsymbol{h_t^b})$を連結した$H(\boldsymbol{h_1}, \boldsymbol{h_2}, \ldots \boldsymbol{h_n})$が Attention 層に入力され（$n = 2t$），Self-Attention $A = softmax(W_{s2} tanh(W_{s1} H^T))$を計算する．$t$は入力の系列長であり，入力したファイルの単語数によって可変である．W_{s1}, W_{s2}はともにパラメータで，$(d_a, 2N_L)$型の行列である．d_aはハイパーパラメータであり，$d_a = 50$とした．Query，Key，Value にはすべてHが渡される．HとAの内積$M(\boldsymbol{m_1}, \boldsymbol{m_2}, \ldots, \boldsymbol{m_n})$の総和$\sum_{i=1}^{n} \boldsymbol{m_i}$を Attention 層の出力として，Dense 層に入力する．

表 20 に学習に使用したデータの内訳を示す. 表 21 に，(a).o，(b).gunk，(c).ar，(d).c，(e).ipi を未知亜種検証データとしたときの結果を示した．既知亜種を F 値 0.9793 で，未知亜種を F 値 0.9950 で検知した．

9.2 実験 3.b – ファミリ分類

BiLSTM と Attention モデルをベースとした推定モデルを設計し，良性通信と 5 クラスのマルウェアファミリについて分類精度を評価する．モデルの構造は 9.1 項と同じであるが，$N_{D2} = 6$，活性化関数をソフトマックス関数として，良性通信と 5 種類のマルウェアファミリに分類する．

表 5 に，学習に使用したデータの内訳を示す．20 epoch の学習を行った結果を，表 22 から表 25 に示した．実験 1.b で見られたような, 特定のクラスに偏って分類される傾向は弱くなった．表 9 と表 25 を比較すると訓練データの分類精度は LSTM よりも高くなり，既知の亜種と未知の亜種の分類精度にわずかな向上が見られた．

10. まとめ

パケットのヘッダ情報から「単語」を定義して，通信トラフィックの分散表現学習を基にしたマルウェア検知モデルを構築し，良性通信・悪性通信の分類ならびにファミリ分類を行った．

良性通信・悪性通信の分類実験では，訓練データに含まれる既知のマルウェアに加え，未知の亜種を水野らの先行研究よりも高精度に検知できている．ただし，LSTM によるモデルでは実際には良性である通信を悪性通信であると誤検知する場合があり，BiLSTM と Attention によるモデルでは実際には悪性である通信を良性通信であると誤検知する場合が見られた．

ファミリ分類実験では，LSTM によるモデルでは特定のファミリに偏って分類され，検知モデルがマルウェアファミリごとの通信の特徴をうまく学習できていなかった．BiLSTM と Attention モデルでは訓練データについて高い性能を示したが，汎化性能が低く未知の亜種の検知性能も低い結果となった．加法構成性を有する分散表現を用いることなどによって，高精度なファミリ分類を行う推定モデルの構築が今後の課題である．

参考文献

[1] 水野翔, 畑田充弘, 森達哉, 後藤滋樹. HTTP ヘッダフィールドの可変性に基づくマルウェア感染端末の特定. コンピュータセキュリティシンポジウム 2016 論文集, 2016(2), pp. 632-639.
[2] 武部嵩礼. 動的解析の Deep Learning による亜種マルウェア推定法 Master's thesis, 早稲田大学基幹理工学研究科情報理工・情報通信専攻, 2016.
[3] 西尾泰和. word2vec による自然言語処理. オライリー・ジャパン, 2017.
[4] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research, Vol. 11, No. Feb, pp. 625-660, 2010.
[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pp. 3111{3119, 2013.
[7] Jacob Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
[8] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In International Conference on Machine Learning, pp. 1188{1196, 2014.
[9] “The Malrec Dataset”. https://giantpanda.gtisc.gatech.edu/malrec/dataset/
[10] Giorgio Severi, Tim Leek, and Brendan Dolan-Gavitt. "Malrec: compact full-trace malware recording for retrospective deep analysis." International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Springer, Cham, 2018.
[11] Sebastián M., Rivera R., Kotzias P., Caballero J. (2016) AVclass: A Tool for Massive Malware Labeling. In: Monrose F., Dacier M., Blanc G., Garcia-Alfaro J. (eds) Research in Attacks, Intrusions, and Defenses. RAID 2016. Lecture Notes in Computer Science, vol 9854. Springer, Cham. https://doi.org/10.1007/978-3-319-45719-2_11
[12] “Kaspersky”. https://www.kaspersky.co.jp/
[13] “Rules for naming | encyclopedia by Kaspersky”. https://encyclopedia.kaspersky.com/knowledge/rules-for-naming/
[14] “VirusTotal”. https://www.virustotal.com/gui/
[15] “tshark”. https://www.wireshark.org/docs/man-pages/tshark.html
[16] 寺田真敏, 他: マルウェア対策のための研究用データセット MWS Datasets ～コミュニティへの貢献とその課題～. 情報処理学会, Vol.2020-IFAT-139 No.8, 2020 年 7 月.
[17] Stratosphere Lab, “Normal Datasets”. https://www.stratosphereips.org/datasets-normal
[18] “Keras: the Python deep learning API”. https://keras.io/
[19] Lin, Zhouhan, et al. "A structured self-attentive sentence embedding." arXiv preprint arXiv:1703.03130 (2017).