7X-06

深層学習を用いたTorブラウザのアクセス識別における有効な特徴点の分析

木檜　圭祐 †i　　利光　能直 ‡　　北條　大和 ‡　　菊田　翼 ‡　　高山　眞樹 ‡
儀員　竜真 †　　齋藤　孝道 †
明治大学 †　明治大学大学院 ‡

1　はじめに

インターネットで通信する際，ネットワークの接続経路を匿名化する The Onion Router(Tor) と呼ばれるシステムがある．また，Tor ブラウザは Tor を使って匿名通信が可能な Web ブラウザである．Tor の追跡に関する研究として，利光ら [1] は，Tor ブラウザで Web サーバに接続した際のアクセスデータを基に，2 つのサーバへのアクセスが同じ Tor ブラウザからのアクセスかどうかの識別 (以降，識別と呼ぶ) が可能であることを示した．しかし，アクセスデータの内，どの情報が識別に有効かが明らかでなかった．本論文ではランダムフォレストを用いてアクセスデータの中で識別に有効な情報を調査した．

2　関連技術

2.1　The Onion Router(Tor)

Tor は，ネットワークの接続経路を匿名化できるシステムである．Tor を使った通信は，Onion Router(OR) と呼ばれるプロキシを 3 つ経由する．この仕組みによって，接続先に自分の IP アドレスを知られることなく通信をすることができる．

3　実験の方法

3.1　実験データ

本論文では，2019 年 8 月 22 日から 2019 年 11 月 28 日の間に収集した，152 端末，総数 3,889 個の Tor ブラウザからのアクセスデータを実験に利用した．アクセスデータは，どの端末からのアクセスかを識別するための識別子を入力するフォーム欄を設置した実験用 Web サイトを用意し，定期的にアクセスしてもらうことで収集した．

3.1.1　ベクトルデータの作成

実験のために，収集したアクセスデータをベクトルデータに変換する．ランダムに 2 件のアクセスデータを抽出し，結合した組を作成し組み合わせとする．

Consideration on required skill set based on Security Incidents
†Keisuke KOGURE　‡Yoshinao TOSHIMITSU　‡Yamato HOJYO
‡Tsubasa KIKUTA　‡Masaki TAKAYAMA　†Tatsuma ISOGAI
†Takamichi SAITO
†Meiji University
‡Graduate School of Meiji University

3.2　実験 1 ランダムフォレストによる実験

実験 1 では，ランダムフォレストを利用し，識別において有効な，端末の特徴となる情報（以降，特徴点と呼ぶ）を調査した．初めに，アクセスデータをそのまま使うとランダムフォレストが行えないため，アクセスデータを 103 個の特徴点に分類した．そして，scikit-learn の RandomForestClassifier を使用して決定木探索を行い，アクセスデータの識別において有効な特徴点を調査した．図 1 に実験 1 の概要を示す．

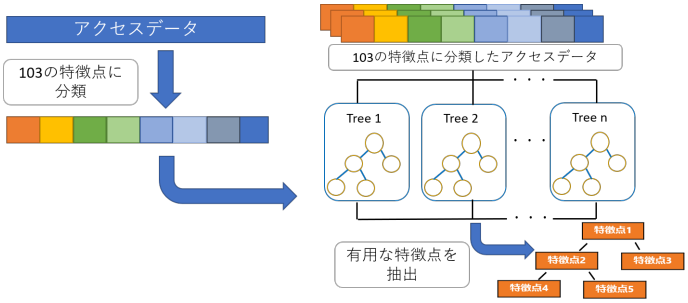

図 1: 実験 1 の概要

3.3　実験 2 深層学習による実験

実験 2 では，3.2 で識別に有効であることが確認できた特徴点のみを利用して深層学習を行い，識別をした．また，比較として 3.2 で作成した 103 個の特徴点をすべて利用した場合でも同様に識別を行った．

深層学習を行うために 3.1.1 の手順で作成したベクトルデータから学習用データ，検証用データ，およびテストデータを作成する．ベクトルデータをランダムに 8:2 に分け，学習用データ，テストデータとそれぞれし，さらに作成した学習用データを 8:2 に分け，学習用データと検証用データとそれぞれする．

3.3.1　ニューラルネットワークの構造

本論文において，ニューラルネットワークは中間層 4 層で構成し，各層で Batch Normalization を行い，活性化関数は relu，損失関数は binary crossentropy，最適化関数には Adagrad（パラメータは Keras のデフォルトの値）を用いた．

4 実験結果

4.1 識別精度算出に使用した指標

モデルの評価では Precision，Recall，Specificity，Accuracy，F1 値を使用する．

4.2 実験 1 について

実験 1 で使用した特徴点 103 個の内，51 個が識別に有効であると判断された．識別において特に有効であった上位 10 個の特徴点を表 1 に示す． 今回の実験では

表 1: 識別において有効な特徴点 (上位 10 個)

rank	特徴点名
1	Form_item （スクリーンサイズ，メモリサイズ，cpu コア数など）
2	Value
3	Internet Protocol Version，送信元 ip，送信先 ip
4	送信元 ip
5	Key
6	送信元ポート番号
7	Stream index
8	iRTT
9	Bytes sent since last PSH flag
10	Time to live

協力者に識別子を送信してもらう際にスクリーンサイズ，メモリサイズ，cpu コア数などの端末の特徴となる値を併せて採取しており，その値を Form_item として使用している．表 1 の Key，Value は Form_item と同じものを表している．Key にはスクリーンサイズ，メモリサイズなどの項目が，Value には具体的な値が入っており，Form_item には Key と Value の 2 つを合わせたものになっている．表 1 の 3，4 番目は送信元 ip に関する情報であった．

以上のことから，Form_item に入っているスクリーンサイズ，メモリサイズなどの特徴点と送信元 ip が特にアクセスデータの識別に有効な特徴点とされていることが分かった．

4.3 実験 2 について

表中の数値は，特に断りがない限り小数点第 4 位を四捨五入した数値である．103 個の特徴点全てを使用した識別の結果を表 2 に示す．また，4.2 で有効だと判断された特徴点 51 個を使用した識別の結果を表 3 に示す．

表 2: すべての特徴点を用いた場合の結果

Pre	Rec	Acc	Spe	F1
0.933	0.784	0.852	0.996	0.982

表 3: 有効な特徴点 51 個を用いた場合の結果

Pre	Rec	Acc	Spe	F1
0.994	0.972	0.983	1.000	0.998

表 2，表 3 から，4.2 で有効だと判断された特徴点 51 個を学習に使用することで，全ての指標について，103 個の特徴点全てを学習に使用した場合と比較して精度が向上し，0.9 以上と高い精度となった．

5 考察

5.1 識別を高い精度で行えたことについて

Form_item は端末の特徴となる値であり，アクセスデータ中に多く含まれていたので，高い精度で識別が可能であったと考えられる．

5.2 精度の向上について

表 2，表 3 から識別に有効でない特徴点がアクセスデータ内に多く含まれていることが分かった．そして，実験 2 で使用した特徴点は 4.2 で少しでも有効であるとされたものを全て使用した．そこで，有効である特徴点 53 個の中から有効度が低い情報をさらに削除し，識別に最適な特徴点の組み合わせを検証することで，精度が向上する可能性がある．

6 研究倫理

我々は，Menlo report[2] の精神に則り，倫理的配慮をして実験を行った．実験を行う際，個人識別はせず，プライバシーを遵守した．本論文で使用されたデータセットの提供元はデータセットの利用目的を理解している．また，研究に使用されたデータセットは，学術的な目的にのみ使用し，我々の研究室にて厳重に保管されており，他者への売却および提供をしない．

7 まとめ

本論文では，ランダムフォレストを用いることで，Tor ブラウザで Web サーバに接続した際のアクセスデータの中でどの特徴点が識別に有効か調査した．そして，有効だと判断された特徴点 51 個を用いることで深層学習による識別をより高い精度で行えることを示した．また，アクセスデータの内，スクリーンサイズ，メモリサイズなどの端末の特徴となる情報と送信元 ip が特に識別に有効であることが分かった．

参考文献

[1] 利光 能直，齋藤 祐太，北條 大和，野田 隆文，齋藤 孝道，深層学習を用いた Tor ブラウザアクセス識別の試み，2020 暗号と情報セキュリティシンポジウム（2020）

[2] Dittrich, D. and Kenneally, E. The Menlo Report:Ethical Principles Guiding Information and Communication Technology Research. U.S. Department of Homeland Security, Aug 2012.