様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 １８ 日現在

機関番号：３２６９２
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１２４８

研究課題名（和文）機械学習による誤りが引き起こす情報セキュリティ問題に関する研究

研究課題名（英文）Research for Problems in Information Security Caused by Application of Machine Learning

研究代表者
宇田　隆哉（Uda, Ryuya）

東京工科大学・コンピュータサイエンス学部・准教授

研究者番号：５０３５０５０９

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：もっとも大きな成果は、画像の見た目が大きく変化しないで強い除去耐性を持たせたAdversarial CAPTCHAの開発である。その他、クロス・サイト・スクリプティングを機械学習により検出する手法の問題点を指摘する研究、人間に判読が困難なナンバープレートを機械学習で読む研究、筆記の特徴を分解することで、任意の文字の筆記に対応した本人確認の研究、マルウェアから特徴を抽出してサイズを縮小することで、機械学習の時間を削減した上で高精度なマルウェア検出を行える研究も行った。

研究成果の学術的意義や社会的意義
機械学習があらゆるものに利用されるようになり、情報セキュリティ分野のサービスやシステムにも利用されるようになってきた。一方、機械学習に詳しい者が情報セキュリティに詳しいとは限らず、また逆もしかりであるため、開発された技術に問題がある場合や、開発自体を断念してしまうこともあり得る。研究成果のAdversarial CAPTCHAは、万能と思われていた人工知能技術に一石を投じるものであったと言える。XSS検出技術における問題点の指摘や、大量の写真を使わずにナンバープレートの数字を読む技術は、通常とは異なる視点からの解を社会に与えられたと考えている。

研究成果の概要（英文）：The greatest effort is Adversarial CAPTCHA which has strong removal resistance while keeping its visibility. Other researches are pointing out a problem in researches for detecting Cross Site Scripting with machine learning, recognizing unreadable numbers on license plates by machine learning, personal verification with any hand-written scripts by disassembling writing features and high accuracy malware detection by size compression with malicious features while reducing time for machine learning.

研究分野：情報セキュリティ

キーワード：情報セキュリティ　人工知能　機械学習

１．研究開始当初の背景

本研究は、機械学習を使用したシステムに対して、機械学習によって引き起こされる問題点を明らかにすることで、情報セキュリティを専門としない技術者にシステムの改善を促すことを目的としていた。機械学習の台頭により人工知能の技術が飛躍的に向上し、新しいサービスやシステムが続々と誕生した。さらに、IoT 時代を迎え、様々な機器がインターネットを通して相互に接続されるようになった。しかし、機械学習には様々な問題点があり、使い方を誤ると有益ではなく有害となる。情報セキュリティを専門としない技術者が開発したものの中に、機械学習における問題点を含むものがあれば、インターネットを通してそれらが攻撃を受けてしまう。そこで、本研究では、それらの問題点を調査し発見することを考えた。本研究の成果を知らしめることで、機械学習を使ったサービスやシステムから脆弱性を除去できるというのが研究開始当初の背景である。

２．研究の目的

各所で機械学習を用いた技術の研究や開発が急速に進む中で、機械学習の仕組みや情報セキュリティの仕組みについて明るくない研究機関や企業が、情報セキュリティと関係のある何かのシステムやサービスに機械学習を適用しようとしており、その問題点や危険性については広く語られていなかった。

例えば、既存システムでは 1%の見逃しがあるものが、機械学習を使用した新システムでは 0.1%になったとする。割合だけで見ると新システムのほうが高性能であるが、この 0.1%の中には、既存システムでは絶対に見逃すことのない攻撃が含まれている可能性がある。情報セキュリティにおいて、この可能性は致命的になることがあると同時に、故意にこの見逃しを再現できた場合にはさらに問題である。

このような問題を見つけて防ぐことが本研究の目的であった。

３．研究の方法

まず、機械学習が誤分類をする際、誤分類されたサンプルの調査を行った。もし、誤分類されたサンプルに共通点があれば、それを他の方法でフィルタすることできる可能性がある。逆に言えば、誤分類されるサンプルが影響を無視できるほど僅かであったとしても、誤分類されるサンプルの性質が明らかになってしまえば、故意に誤分類されるサンプルを量産して攻撃を行うことが可能となる。これは看過できない問題である。

次に、機械学習の演算の仕組みを利用して誤分類を起こさせることができるかどうかについても調査を行った。また、現実世界の機械学習においては、金銭的、時間的な問題から理想的なトレーニングを行うことは困難であることにも気づいた。トレーニングを簡略化するため、関連技術では様々な工夫が用いられているが、そこに情報セキュリティについての問題が起きていることもあった。このような問題が起きる研究やシステムがないか調査し、我々の解決策を提示した。

４．研究成果

(1) もっとも大きな成果は、共同研究者の柴田准教授と開発した Adversarial CAPTCHA という技術である。人間に判読できてコンピュータに判読できない文字列を示すことで、人間とコンピュータを区別する文字列 CAPTCHA は従来より使用されていた。そして、機械学習による人工知能の登場により、これらはすべて人工知能に判別可能とされる研究結果が発表された。一方で、機械学習に誤分類を引き起こさせる Adversarial Examples という技術が注目を集め、画像に Adversarial Examples による微小な変化を加えるだけで、人工知能が画像を誤分類することが分かった。しかし、Adversarial Examples を除去する対策がなされ、除去されないようにするためには見た目の画像が大きく変化するほど強く Adversarial Examples を乗せる必要があった。とくに、CAPTCHA の場合には、これでは人間にも判読不可能になってしまい、意味を成さないという問題があった。これに対して我々の Adversarial CAPTCHA は、摂動幅を小さく

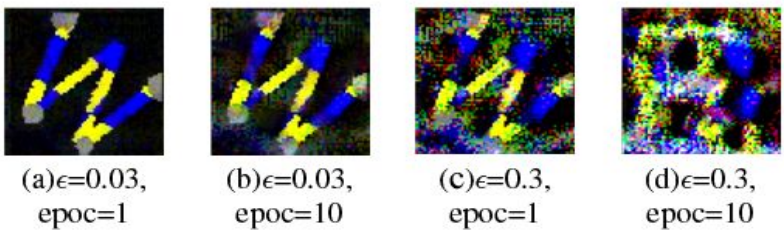

図 1: 文字列 CAPTCHA に対する Adversarial Examples の摂動幅とエポック数

することでエポックを重ねても見た目の影響が少ないことを利用し、画像の見た目がそれほど大きく変化しないで人工知能に誤分類を起こさせるものである。図 1 に示すように、摂動幅が 0.3 の場合にはエポック数が 1 の(c)でも画像が汚く、10 の(d)ではほとんど文字が視認できない状態になっているが、摂動幅が 0.03 の場合にはエポック数が 10 の(b)でも画像の劣化が(c)よりも少ない。画像を作成する具体的な原理の詳細は国際会議にて発表した[1]が、英文ジャーナル論文[2]と和文ジャーナル論文[3]にこの CAPTCHA のシステムが完成するまでの過程としてそれぞれ採録になっている。

(2) 機械学習でクロス・サイト・スクリプティング(XSS)を見分ける研究の問題点を指摘する研究も行った[4]。既存研究において、機械学習を使用すれば XSS を高精度で検出できるとするものが複数存在した。しかし、我々が確認した時点で、すべての既存研究のデータセットには問題があり、それらの手法で実際に攻撃可能な XSS を検出しようとすると、非常に精度が低くなることが判明した。一方、機械学習の前処理に工夫を施した我々の手法で XSS の検出を行えば、検出精度は下がらない。なお、実際の攻撃に使用された XSS のサンプルを多数集めることは非常に困難であるため、機械学習がうまく行えないことを示す擬似データセットを開発する研究も行った[5]。これは、この擬似データセットを高精度で分類できれば実検体も高精度で分類できることを保証するものではなく、この擬似データセットを高精度で分類できないにも関わらず実検体を高精度で分類できている場合には、使用しているデータセットに問題があることを示すものである。本研究は優秀研究賞を受賞した。

(3) 人間に判読が困難なナンバープレートを機械学習で読むという研究も行った[6]。通常、このような場合には、人間に判読が困難なナンバープレートの画像を大量に用意する必要があり、不可能ではないが困難である。本研究においては、日本のナンバープレートのフォントと大きさが一意であることに着目し、これを CG で再現することで、トレーニング用の実画像を 1 枚も用意することなく高精度で判別が可能となった。この成果は和文ジャーナル論文に掲載された。

(4) 筆記の特徴を分解することで、任意の文字の筆記に対応した本人確認の研究も行った[7]。筆記の特徴を機械学習することは従来より可能であるが、署名であれば個別に学習が必要であり、漏洩した場合に署名を変更できない問題がある。本研究では、分解した筆記の特徴を学習することで、任意の文字列を本人確認に使用できる。これは、機械学習を前提とした署名の漏洩に対して非常に有効である。この成果は和文ジャーナル論文に掲載された。

(5) マルウェアから特徴を抽出してサイズを縮小することで、機械学習の時間を削減した上で高精度なマルウェア検出を行える研究も行った[8]。従来も、マルウェアを機械学習で検出する研究は行われていたが、オリジナルのサイズで機械学習を行うことはコスト的に現実的ではなかった。そこで、マルウェアから特徴を抽出し、その特徴を機械学習する手法が多く提案されたが、それらの手法には検出を回避可能な問題点が存在するものや、数パーセント程度の見逃しを許容するものがあった。これに対して、本研究では、亜種マルウェアに限定すれば完全な検出が行えることを示した。この成果は和文ジャーナル論文に掲載された。

＜引用文献＞
[1] Tomoka Azakami, Chihiro Shibata, Ryuya Uda and Toshiyuki Kinoshita, Creation of Adversarial Examples with Keeping High Visual Performance, Proceedings of the IEEE 2nd International Conference on Information and Computer Technologies, pp.52-56, 2019.
[2] Tomoka Azakami, Chihiro Shibata and Ryuya Uda, Evaluation of Ergonomically Designed CAPTCHAs using Deep Learning Technology, Journal of Information Processing, Vol.59, No.9, 2018.
[3] 阿座上知香, 柴田千尋, 宇田隆哉, Adversarial CAPTCHA：畳込みニューラルネットワークに耐性のある CAPTCHA の提案と評価, 情報処理学会論文誌, Vol.60, No.2, pp.680-695, 2019.
[4] Sota Akaishi and Ryuya Uda, Classification of XSS Attacks by Machine Learning with Frequency of Appearance and Co-occurrence, Proceedings of the 53rd Annual Conference on Information Sciences and Systems (CISS), 2019.
[5] 飯野和真, 宇田隆哉, 不適切なデータセットや処理方法を用いた機械学習による XSS 攻撃検出研究の解説と精度の比較, 情報処理学会研究報告, CSEC-92, Vol.2021-CSEC-92,No.20,1-8
[6] 鈴木友哉, 宇田隆哉, CNN を用いた予測に有効なナンバープレート写真用トレーニングデータの検討, 情報処理学会論文誌, Vol.62, No.2, pp.484-496, 2021.
[7] 釜石智史, 宇田隆哉, 特徴の再訓練を必要としない変更可能な筆記, 情報処理学会論文誌, Vol.63, No.4, pp.1094-1114, 2022.
[8] 瀧口翔貴, 宇田隆哉, n-gram 抽出と機械学習を用いた亜種マルウェア分類手法の提案と評価, 情報処理学会論文誌, Vol.63, No.4, pp.1052-1071, 2022.

５．主な発表論文等

〔雑誌論文〕　計13件（うち査読付論文　8件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 瀧口翔貴，宇田隆哉	４．巻 63
２．論文標題 n-gram抽出と機械学習を用いた亜種マルウェア分類手法の提案と評価	５．発行年 2022年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 1052-1071
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 釜石智史，宇田隆哉	４．巻 63
２．論文標題 特徴の再訓練を必要としない変更可能な筆記	５．発行年 2022年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 1094-1114
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 宇田隆哉	４．巻 1
２．論文標題 n-gramによるマルウェア検出における機械学習を騙す良性ソフトウェア汚染	５．発行年 2021年
３．雑誌名 コンピュータセキュリティシンポジウム2021論文集	６．最初と最後の頁 623-630
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 宇田隆哉	４．巻 なし
２．論文標題 圧縮サイズと比較コストを考慮したマルチN-gramによる亜種マルウェアの検出	５．発行年 2020年
３．雑誌名 情報処理学会コンピュータセキュリティシンポジウム2020論文集	６．最初と最後の頁 不明
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 飯野和真，宇田隆哉	4．巻 2021-CSEC-92
2．論文標題 不適切なデータセットや処理方法を用いた機械学習によるXSS攻撃検出研究の解説と精度の比較	5．発行年 2021年
3．雑誌名 情報処理学会研究報告	6．最初と最後の頁 1-8
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Masaki Shiraishi and Ryuya Uda	4．巻 なし
2．論文標題 Comparison of Algorithms and Action Coordinates Sets in Detection of Slight Differences in Motions like Lock-Picking	5．発行年 2020年
3．雑誌名 Proceeding of the 5th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference	6．最初と最後の頁 なし
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Kouhei Kita and Ryuya Uda	4．巻 なし
2．論文標題 Malware Subspecies Detection Method by Suffix Arrays and Machine Learning	5．発行年 2021年
3．雑誌名 Proceeding of the 55th Annual Conference on Information Sciences and Systems	6．最初と最後の頁 なし
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Tomoya Suzuki and Ryuya Uda	4．巻 なし
2．論文標題 Recognition of Digits on License Plate by RAISR with Changing Contrast Ratio	5．発行年 2021年
3．雑誌名 Proceeding of the 55th Annual Conference on Information Sciences and Systems	6．最初と最後の頁 なし
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 宇田隆哉	4．巻 1
2．論文標題 N-gram抽出法による亜種マルウェアの検出と攻撃耐性の考察	5．発行年 2019年
3．雑誌名 情報処理学会コンピュータセキュリティシンポジウム2019論文集	6．最初と最後の頁 515-522
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 阿座上知香，柴田千尋，宇田隆哉	4．巻 60
2．論文標題 畳込みニューラルネットワークに耐性のあるCAPTCHAの提案と評価	5．発行年 2019年
3．雑誌名 情報処理学会論文誌	6．最初と最後の頁 680-695
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 瀧口翔貴，宇田隆哉	4．巻 118
2．論文標題 N-gram圧縮と深層学習を用いたマルウェア分類手法の提案	5．発行年 2019年
3．雑誌名 電子情報通信学会技術研究報告	6．最初と最後の頁 111-116
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Sota Akaishi and Ryuya Uda	4．巻 なし
2．論文標題 Classification of XSS Attacks by Machine Learning with Frequency of Appearance and Co-Occurrence	5．発行年 2019年
3．雑誌名 The 53rd Annual Conference on Information Sciences and Systems	6．最初と最後の頁 1-6
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/CISS.2019.8693047	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Tomoka Azakami, Chihiro Shibata, Ryuya Uda and Toshiyuki Kinoshita	4．巻 なし
2．論文標題 Creation of Adversarial Examples with Keeping High Visual Performance	5．発行年 2019年
3．雑誌名 IEEE 2nd International Conference on Information and Computer Technologies	6．最初と最後の頁 52-56
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/INFOCT.2019.8710918	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計10件（うち招待講演　0件／うち国際学会　5件）

1．発表者名 宇田隆哉
2．発表標題 n-gramによるマルウェア検出における機械学習を騙す良性ソフトウェア汚染
3．学会等名 情報処理学会，コンピュータセキュリティシンポジウム2021
4．発表年 2021年

1．発表者名 宇田隆哉
2．発表標題 圧縮サイズと比較コストを考慮したマルチN-gramによる亜種マルウェアの検出
3．学会等名 情報処理学会コンピュータセキュリティシンポジウム2020
4．発表年 2020年

1．発表者名 飯野和真，宇田隆哉
2．発表標題 不適切なデータセットや処理方法を用いた機械学習によるXSS攻撃検出研究の解説と精度の比較
3．学会等名 情報処理学会コンピュータセキュリティ研究会
4．発表年 2021年

１．発表者名
Masaki Shiraishi and Ryuya Uda

２．発表標題
Comparison of Algorithms and Action Coordinates Sets in Detection of Slight Differences in Motions like Lock-Picking

３．学会等名
2nd International Workshop on Security and Reliability of IoT Systems（国際学会）

４．発表年
2020年

１．発表者名
Kouhei Kita and Ryuya Uda

２．発表標題
Malware Subspecies Detection Method by Suffix Arrays and Machine Learning

３．学会等名
55th Annual Conference on Information Sciences and Systems（国際学会）

４．発表年
2021年

１．発表者名
Tomoya Suzuki and Ryuya Uda

２．発表標題
Recognition of Digits on License Plate by RAISR with Changing Contrast Ratio

３．学会等名
55th Annual Conference on Information Sciences and Systems（国際学会）

４．発表年
2021年

１．発表者名
宇田隆哉

２．発表標題
N-gram抽出法による亜種マルウェアの検出と攻撃耐性の考察

３．学会等名
情報処理学会コンピュータセキュリティシンポジウム2019

４．発表年
2019年

１．発表者名
瀧口翔貴，宇田隆哉

２．発表標題
N-gram圧縮と深層学習を用いたマルウェア分類手法の提案

３．学会等名
電子情報通信学会

４．発表年
2019年

１．発表者名
Sota Akaishi and Ryuya Uda

２．発表標題
Classification of XSS Attacks by Machine Learning with Frequency of Appearance and Co-Occurrence

３．学会等名
IEEE（国際学会）

４．発表年
2019年

１．発表者名
Tomoka Azakami, Chihiro Shibata, Ryuya Uda and Toshiyuki Kinoshita

２．発表標題
Creation of Adversarial Examples with Keeping High Visual Performance

３．学会等名
IEEE（国際学会）

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	柴田　千尋 (Shibata Chihiro) (00633299)	法政大学・理工学部・准教授 (32675)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関