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 １２ 日現在

機関番号：３２６７８
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ１１９５０

研究課題名（和文）暗号化された複雑なWeb通信のPassive計測によるサービスタイプ特定の研究

研究課題名（英文）Study on identifying HTTP communication service type over encrypted transport layer

研究代表者
塩本　公平（Kohei, Shiomoto）

東京都市大学・情報工学部・教授

研究者番号：００５３５７５０

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：機械学習モデルを用いてサービス特定を行う際に，暗号化されたパケットキャプチャデータから抽出できる特徴量のうち，コネクション毎のバイト数とパケット数に関して最大値，平均値，中間値，および分散がサービス特定に有効な特徴量であることを明らかにした．
半教師あり学習である敵対的自己符号化器を用いたネットワーク侵入検知システムを検討し，訓練データセット中のラベル付きデータサンプルの0.1%のみで，多層パーセプトロンベースのネットワーク侵入検知システムと同等の性能を達成でき，人手でのラベル付け作業が必要となる教師データの数を抑えることができることを明らかにした．

研究成果の学術的意義や社会的意義
今日のWeb通信は複雑な構造であり，暗号化されたパケットキャプチャデータからWebサービスを特定することは困難であった．本研究の学術的意義は暗号化されたパケットキャプチャデータを機械学習によりサービス特定を行う際に有効な特徴量を明らかにしたこと，半教師あり学習を用いることで人手のかかるデータへのラベル付け作業を削減したことである．
本研究の社会的意義は，暗号化が普及した現在のインターネットで困難であったWebサービスの特定が可能になり，ユーザに提供するサービス体感品質を最適化するために必要なネットワーク性能をネットワーク事業者が把握することが可能となったことである．

研究成果の概要（英文）：We analyzed measured data of Web communications consisting of a large number of encrypted connections, and investigated a method for identifying the service types used by users. We found that, among the features that can be extracted from the encrypted packet capture data, the maximum, mean, median, and variance with respect to the number of bytes and packets per connection are the most effective features for service identification using machine learning models.

A semi-supervised learning, adversarial self-coder-based network intrusion detection system is proposed and its performance is evaluated on the NSL-KDD dataset. We found that the proposed method can achieve the same performance as a multi-layer perceptron-based network intrusion detection system with only 0.1% of the labeled data samples in the training dataset, reducing the number of supervised data samples that need to be manually labeled.

研究分野：情報ネットワーク

キーワード：暗号化　Web通信トラフィック　サービスタイプ特定　機械学習　半教師あり学習　特徴量　ラベル付与

様　式　C－１９，F－１９－１，Z－１９（共通）

１．研究開始当初の背景

（１）Web 通信の進化に伴い，ユーザは Web 通信を通じて様々なサービスを利用するようになった．サービスのユーザ体感品質は通信速度や遅延時間などの要因によって決まるが，サービス毎にどの要因が支配的となるかは異なる．例えば，ブラウジングをしている場合は，遅延時間がユーザ体感品質を決める支配的な要因であるが，動画視聴をしている場合は，通信速度が支配的な要因となる．このように，ユーザ体感品質を満足させるためには，ユーザが受けているサービスを特定して，ユーザ体感品質を決める支配的な要因を制御することが必要となる．

（２）近年の暗号化 Web 通信の普及に伴い，ユーザが Web 通信によって受けているサービスを特定することが困難となっている．暗号化された通信から取得したパケットキャプチャデータからサービスを特定する必要があるが，Web 通信の構造は年々複雑さを増しており，暗号化されたパケットキャプチャデータから Web サービスを特定することは困難であった．

２．研究の目的

本研究の目的は，HTTP 通信のコネクションやパケットの複雑な振る舞いを観測し，それらの統計的な特徴から機械学習を応用して，インターネットで提供される広範囲なサービスの中でユーザが利用しているもの特定することである．具体的には（１）有効な特徴量の検討と（２）半教師あり学習を用いた方法の検討に取り組んだ．

（１）有効な特徴量の検討

観測されるセッションやパケットに関する統計量のうち，サービス識別に有効なものについて明らかにする．ユーザが Web ブラウザで操作を行った際に生じる多数のコネクションと各コネクションで送受されるパケットに関してサービス毎に大きく異なる特徴量を見出す．

（２）半教師あり学習を用いた方法の検討

少ないラベルの付いた教師データを用いて学習できる機械学習を用いた識別器の構成法を明らかにする．教師あり学習を用いた識別器は高精度の識別性能を得るためにはラベルの付いた多量のデータが必要である．しかしながら，教師データのラベル付け作業は人手によるものであり，ラベルの付いた多量の教師データを用意することは困難である．そこで，半教師あり学習を用いた識別器の構成法を検討する．

３．研究の方法

（１）有効な特徴量の検討

研究室内のネットワークにおいてパケットをキャプチャしデータセットを作成した．一定時間ごとに PCAP 形式のデータを分割してデータサンプルを生成した．一方，Web 通信を行っているクライアント PC においてブラウザログを収集し，クライアント PC がどのような Web サービスを利用しているかの情報を取得した．PCAP 形式のデータとブラウザログを照合して，データセットを自動で生成するための Python プログラムを作成し，データセットを構築した．

作成したデータセットからパケットレベルおよびコネクションレベルの特徴量の候補を抽出し，どの特徴量の組み合わせが Web サービスを特定するのに有効であるかを評価した．機械学習アルゴリズムとしては Multi-layer Perceptron（MLP）を用いて評価を進めた．データセットを入力すると，特徴量を組み合わせて，MLP によりサービス特定のための学習と推論を行う Python プログラムを作成して実験を進めた．

（２）半教師あり学習を用いた方法の検討

人手のかかるデータへのラベル付け作業を削減するために半教師あり学習を用いた手法を検討する．ネットワーク侵入検知を目的として，半教師あり学習である敵対的自己符号化器（Adversarial Auto-Encoder：AAE）の学習と推論を行う Python プログラムを作成して，公開の NSL-KDD データセット を用いてシミュレーションにより評価した．

４．研究成果

（１）データセットの作成

研究室で通信を集約するスイッチからパケットをキャプチャし，より実際のトラフィックデータに近いデータを収集した．スイッチハブにおいてミラーリングして実験室内の通信をキャプチャした．ブラウジング用の PC は，研究室の他の PC と同様にイーサネットで接続し，ブラウジングには Chrome ブラウザを使用した．パケットキャプチャ用の PC は，スイッチハブのミラーリングポートに接続した．パケットキャプチャしたデータを用いてデータセットを作成し，得られた PCAP データを分割した．図１に示すように，ウィンドウサイズを w としウィ

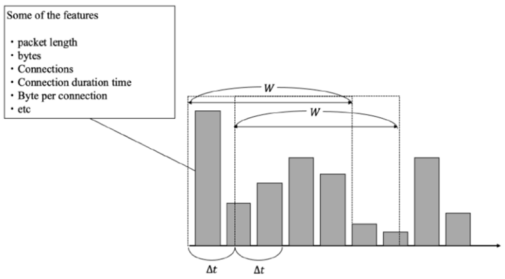

図１　PCAP データからのデータサンプルの抽出

ンドウサイズ内に観測された PCAP データを 1 つのデータサンプルとした．ウィンドウサイズの始点を Δt だけ移動させ，そこから w までのデータを 1 つのデータサンプルとして PCAP データの終点までデータサンプルを収集した．

ブラウザログを利用してデータサンプルにラベル付けを行った．PCAP から HTTP GET メッセージを抽出するだけでは，ユーザのクリックと自動的に誘導された HTTP GET メッセージの区別がつかず，ユーザの利用状況を正しく把握することができない．そこで，ブラウザのログを解析して，クリックされた URL を収集することにした．図 2 に示すように，ブラウザログから求めた URL クリック時刻を PCAP データと照合し，PCAP データから得たデータサンプルにサービス種別のラベルを付与した．ブラウザログからサイト遷移のタイミングを割り出し，PCAP データで同じタイミングを一定の時間間隔で区切ってラベルを付与した．アクセスしたサイトの URL を解析して同一サイトを判定し，各サイトにラベルを付与した．

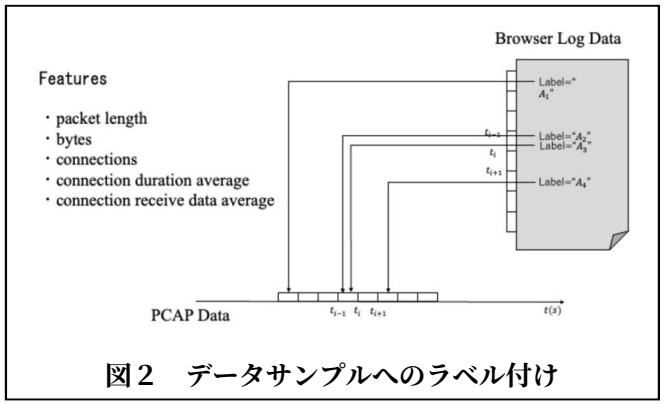

図 2　データサンプルへのラベル付け

データセットは w = 8，Δt = 1 で作成した．各データサンプルはラベルが付与されており，そのうち 25 ラベルがブラウジング用で，2 ラベルがストリーミング用である．ブラウジングにはニュースサイトや辞書サイト等のテキストと画像が中心で動画や音声がほとんどない Web サイトであり，ストリーミングは youtube.com と www.amazon.co.jp である．

（2）有効な特徴量の検討

機械学習モデルの入力となる特徴量を検討した．パケットレベルの特徴として表 1 のものを，コネクションレベル特徴として表 2 のものを検討した．コネクションレベルの特徴では，コネクションを 5 タプルで識別し，ウィンドウサイズ w の間の接続の特徴を得た．識別された各コネクションの継続時間から，PCAP をウィンドウサイズ w で分割したデータセットがどの時点に相当するかを判断し，当該分割時刻に処理されたデータとして扱った．特徴量のうち，どの成分が有効かを特定するために，表 3 の特徴量の組み合わせで評価した．例えば，特徴セット s1 はすべてのパケットレベル特徴とすべての接続レベル特徴を含み，MLP の入力層が 31 次元であることを，特徴セット s2 はすべてのパケットレベルの特徴とホスト数を除くすべてのコネクションレベルの特徴を含むことを意味する．

Type	Statistics	Direction
Number of Hosts	Total	-
Packet Counts	Mean/Variance	fwd/bwd
Bytes per Packet	Mean/Variance	fwd/bwd

表 1　パケットレベルの特徴量

Type	Statistics	Direction
Number of Hosts	Total	-
Number of Connections	Total	-
Number of Connections per Hosts	Mean/Variance	-
Connection holding time	Max/Mean/Median/Variance	-
Packet Counts per Connection	Max/Mean/Median/Variance	fwd/bwd
Bytes per Connection	Max/Mean/Median/Variance	fwd/bwd

表 2　コネクションレベルの特徴量

作成したデータセットを用いて MLP によりサービスを識別した．2601 データサンプルのうち 2000 データサンプルをトレーニングデータとして使用し，残りのデータサンプルをテストに使用した．評価した結果，分類精度は s14 が 91%と最も高く，次いで s3 が 91%，s5 が 90%であった．最も精度が低かったのは s12 で 57%の精度であった．少ない特徴量でも最も精度が高かったのは Packet Counts Per Connection のみを用いた s14 で 88%の精度を示した．s15 は Connection holding time，Packet Counts per Connection，Bytes per Connection の 3 種類の特徴量の組み合わせで 90%の精度

となり，コネクションベースの特徴量のみを用いた組み合わせで最も精度が高かった．以上の結果から，機械学習モデルを用いてサービス特定を行う際に，暗号化されたパケットキャプチャデータから抽出できる特徴量のうち，コネクション毎のバイト数とパケット数に関して最大値，平均値，中間値，および分散がサービス特定に有効な特徴量であることを明らかにした．

Level	Type	$s1$	$s2$	$s3$	$s4$	$s5$	$s6$	$s7$	$s8$
Packet	Packet Counts	√	√	√	√	√	√	√	-
Packet	Bytes per Packet	√	√	√	√	√	√	√	-
Connection	Number of Hosts	√	-	√	-	-	-	-	-
Connection	Number of Connections	√	√	-	-	-	-	-	-
Connection	Number of Connections per Hosts	√	√	-	-	-	-	-	
Connection	Connection holding time	√	√	√	√	√	-	√	-
Connection	Packet Counts per Connection	√	√	√	√	-	-	-	-
Connection	Bytes per Connection	√	√	√	√	√	√	-	√
Level	Type	$S9$	$S10$	$S11$	$S12$	$S13$	$S14$	$S15$	$S16$
Packet	Packet Counts	-	√	-	-	-	-	-	-
Packet	Bytes per Packet	-	√	-	-	-	-	-	-
Connection	Number of Hosts	-	-	√	√	-	-	-	-
Connection	Number of Connections	-	-	√	√	-	-	-	-
Connection	Number of Connections per Hosts	-	-	√	√	√	-	-	-
Connection	Connection holding time	√	-	√	-	√	-	√	-
Connection	Packet Counts per Connection	-	-	√	-	√	√	√	√
Connection	Bytes per Connection	-	-	√	-	√	-	√	√

表３　特徴量の組み合わせパターン

（３）半教師あり学習を用いたネットワーク侵入検知方法の検討

必要なラベル付き訓練サンプル数を減らすために，AAE を用いた半教師付き機械学習によるネットワーク侵入検知システム（Network Intrusion Detection System：NIDS）を提案した．提案手法を一連の実験により評価し，以下の結果を得た：

- 提案した AAE は，ラベル付きデータサンプルの 0.1%だけで，MLP と同等の性能を達成し，少数のラベルなしデータサンプルの追加を行った．
- ラベル付けされたデータサンプルの数が少ない場合，ラベル付けするサンプルの選択に関わらず，精度は変化しなかった．したがって，アノテーションするデータサンプルの選択は，提案する AAE の性能に影響しない．
- データ構造の次元を z2=10 とすることで，データサンプルから本質的な特徴を抽出することに成功し，正常分類と攻撃分類を最も明確に分離し，結果としてリコールと F1 スコアの面で最高の性能を発揮することを実証した．

MLP と AAE のアーキテクチャを図 3 に示す．この例では，AAE に関して，エンコーダは 122 の入力を持ち，重要な特徴に圧縮されて 52（= 2 + 50）の出力（z1 に対して 2，z2 に対して 50）をもたらす．デコーダは，潜在変数ベクトル（z1，z2）を持つ隠れ中間層から 52 の入力を受け，122 の出力を得る．エンコーダとデコーダはともに 1000×1000 の大きさの中間完全連結層を持っている．また，ドロップアウト層があり，層間で一括正規化される．カテゴリ分布の識別器（z1）は 2 つの入力を受け，1 つの出力（"Fake" または "Real"）を得ることができる．ガウス分布の識別器（z2）は 50 個の入力を受け，1 つの出力（「偽物」または「本物」）を出す．カテゴリ分布(z1)とガウス分布(z2)の識別器は，いずれも 1000×1000 の大きさの中間全結合層を持っている．また，層間では一括正規化を行っている．

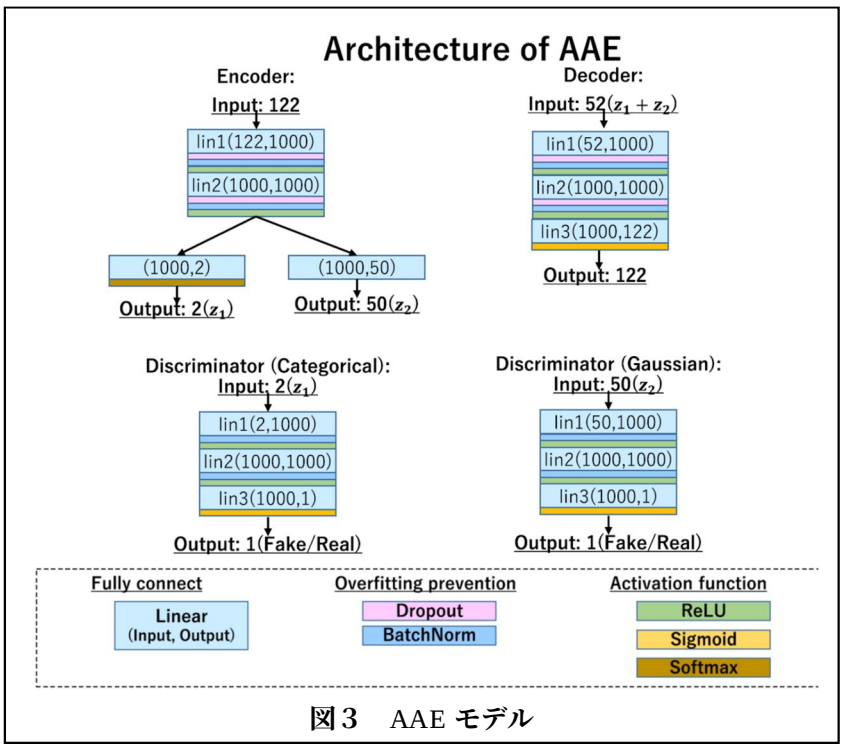

図３　AAE モデル

図 4 は訓練データの 0.1%をラベル付きとして使用した場合の, AAE の精度, 精度，再現性，F1 スコアなどの性能である．横軸はラベルの付いていない訓練サンプルの数を表している．比較のために MLP の性能も示している．図 4 で recall に着目すると，AAE は MLP よりも高い recall を達成していることが確認できる．また，提案する AAE は，訓練データセットに 0.1%のラベル付きデータしかなくても，MLP と同等の性能を達成することが確認できた．次に，「トレーニングデータセットにどれだけのラベルなしデータサンプルが必要なのか」という問題について検討した．図 4 から，ラベルなしデータサンプルを追加することで，AAE の性能が向上することがわかる．このように，AAE は，ラベルなしデータサンプルの分布構造を利用して，ラベル付きデータサンプルから計算される隣接クラス間の境界の精度を向上させることが確認された．

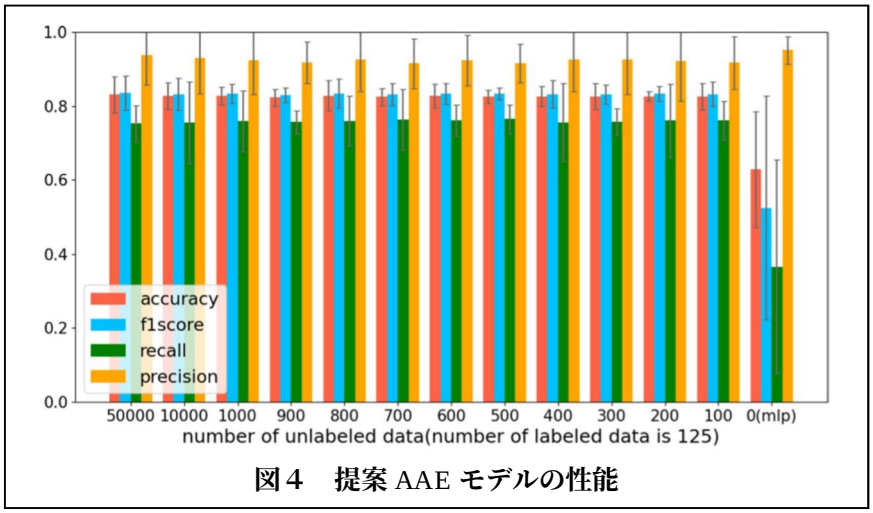

図 4　提案 AAE モデルの性能

（4）研究成果のまとめ

ラベル付きの学習サンプルが少ないために起こる問題を調査した研究は近年注目を集めており, One-shot 学習や Few-shot 学習, Varietional Auto-Encoder（VAE）や Stacked Sparse Auto-Encoder（SSAE）を用いた半教師付き学習、教師なし特徴抽出と教師あり分類アルゴリズムを組み合わせた方法などが提案されており，本研究課題で取り組んだ AAE を用いた半教師あり学習もそのような近年注目を集めている分野に位置づけられる．

機械学習に必要なラベル付きの学習サンプル数を削減できることは，人手によるデータサンプルへのラベル付け作業を削減することができ，機械学習を用いた Web 通信サービスの特定方法の検討を大きく前進させることができる．本研究成果により，暗号化された Web 通信を対象にユーザが利用しているサービスを特定できるようになったことで，ネットワーク運用者がユーザのサービス毎に異なるネットワークの要求性能の実現のための基盤を提供するものである．

この分野の研究は検討の初期段階であり，データ分布や条件に応じて適切なアプローチは適用するための知見を得る必要がある．これまで機械学習を用いた NIDS の性能評価は，ベンチマーク用の公開データセットを用いたものとなっており，実網での有効性は明らかになっていない．このため，機械学習を用いた NIDS はまだ製品化されておらず，実用化に至っていないのが現状である．しかしながら，実網のトラフィックは一般に公開されておらず，実網のトラフィックを用いた有効性の評価を行うことが困難な状況である．このような現状を踏まえて，今後は実トラフィックを計測・収集し，データセットを構築し，機械学習を用いた NIDS の実トラフィックを対象とした有効性を評価していく．

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 吉田 祥太，江口 優，塩本 公平	４．巻 Volume 120, Number 109
２．論文標題 暗号化されたWebサービス推定のためのFew-shot Learningにおけるラベル付与法に関する検討	５．発行年 2021年
３．雑誌名 電子情報通信学会技術研究報告 Online edition: ISSN 2432-6380	６．最初と最後の頁 ICM-37-ICM-42
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 Kazuki Hara and Kohei Shiomoto	４．巻 -
２．論文標題 Intrusion Detection System using Semi-Supervised Learning with Adversarial Autoencoder	５．発行年 2020年
３．雑誌名 Proceedings of IEEE/IFIP Network Operations and Management Symposium (NOMS 2020)	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/NOMS47738.2020.9110343	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 Kohei Shiomoto	４．巻 31
２．論文標題 Network Intrusion Detection System Based on an Adversarial Auto-Encoder with Few Labeled Training Samples	５．発行年 2023年
３．雑誌名 Journal of Network and Systems Management	６．最初と最後の頁 1-22
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10922-022-09698-w	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 原 和希，塩本 公平	４．巻 vol. 118, no. 465
２．論文標題 Adversarial Autoencoderを用いた半教師あり学習によるネットワーク侵入検知システムの検討	５．発行年 2019年
３．雑誌名 電子情報通信学会技術研究報告 Online edition: ISSN 2432-6380	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） 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 －

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　1件）

１．発表者名
江口 優

２．発表標題
インターネットにおける暗号化トラフィックの機械学習に よる分析のための 特徴量と教師データの作成法に関する研 究

３．学会等名
電子情報通信学会　東京支部学生会　第26回研究発表会

４．発表年
2021年

１．発表者名
江口 優

２．発表標題
暗号化されたWebサービス推定のためのFew-shot Learningにおけるラベル付与法に関する検討

３．学会等名
電子情報通信学会　信学技報　2020年7月　ICM研究会

４．発表年
2020年

１．発表者名
Kohei Shiomoto

２．発表標題
Intrusion Detection System using Semi-Supervised Learning with Adversarial Autoencoder

３．学会等名
IEEE/IFIP Network Operations and Management Symposium (NOMS2020)（国際学会）

４．発表年
2020年

１．発表者名
原 和希

２．発表標題
Adversarial Autoencoderを用いた半教師あり学習によるネットワーク侵入検知システムの検討

３．学会等名
電子情報通信学会　信学技報　2019年3月　ＮＳ研究会

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
研究協力者	原　和希 (Hara Kazuki)		
研究協力者	吉田　祥太 (Yoshida Syouta)		
研究協力者	江口　優 (Eguchi Yutaka)		
研究協力者	櫻井　裕生 (Sakurai Yuusei)		
研究協力者	小山　知晃 (Koyama Tomoaki)		

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