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　 ９ 月 ２０ 日現在

機関番号：１２１０２
研究種目：基盤研究(B)（一般）
研究期間：2019～2022
課題番号：１９Ｈ０４１６４

研究課題名（和文）機械と人間の認識ギャップを考慮した深層学習セキュリティ・プライバシーに関する研究

研究課題名（英文）Deep learning security and privacy focused on human-machine recognition gap

研究代表者
佐久間　淳（Sakuma, Jun）

筑波大学・システム情報系・教授

研究者番号：９０３７６９６３

交付決定額（研究期間全体）：（直接経費）　　13,300,000円

研究成果の概要（和文）：AIへの攻撃、AIへの防御、および説明可能なAIの分野において、成果を上げた。代表的な成果は以下の通り。AIへの攻撃では、音声認識モデルを物理世界で攻撃するための敵対的な音声例生成手法を提案した。研究成果はIJCAI2019に採択され、2023年現在引用数は170を超える。AIの防御では、深層学習を使用したコンテンツベース画像検索において、敵対的事例に対する保証付き防御手法を開発した。AI説明性の観点では、深層学習分類器を対象として、データXがクラスYに分類されるのは、XがA、B、を持ち、Cを持たないからである、というタイプの説明を与える手法を提案した。研究成果はAAAI2022に採択された。

研究成果の学術的意義や社会的意義
深層学習が社会にとって重要な判断や意思決定の一部を担うようになった場合，深層学習そのものを不正利用したり，深層学習の判断や意思決定を不正に捻じ曲げて，不当に利益を得ようとする人間が現れると考えられる。そのような敵対的環境において深層学習を適切に動作させるためには深層学習特有のセキュリティの問題を解決する技術が必要である。また深層学習は、学習のために大量にデータを収集したり、予測のために対象に関するデータを取得したりする必要がある。研究ではこのような深層学習のセキュリティに関する問題に対する一定の解決のための方法論を構築した。

研究成果の概要（英文）：Achievements were made in the areas of attacks on AI, defense of AI, and explainable AI. Major results are as follows.In Attacks on AI, we proposed an adversarial audio example generation methodology for attacking speech recognition models in the physical world. The research results were accepted to IJCAI 2019 and have over 170 citations as of 2023. In AI defense, we developed a certified defense methodology against adversarial examples in content-based image retrieval using deep learning. In explainable AI, we proposed a methodology for deep learning classifiers that provides a type of explanation for why data X is classified into class Y because X has A, B, and does not have C. The research results were accepted by AAAI2022.

研究分野：機械学習

キーワード：機械学習　人工知能　セキュリティ　プライバシー　高信頼AI

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

大量データを人間にとって有用な知識やモデルに変換する機械学習は，社会で使われる様々なサービスの中心的な役割を果たすようになりつつある．
従来の機械学習の入力(特徴)は構造情報(主にベクトルや行列)を想定しており，画像や音声からの特徴抽出は 人手による設計が主であった。一方、2010 年代前半において深層学習関連技術が急速に進展し，特に画像や音声を対象とした特徴の自動学習が発展し、深層学習モデルによる認識精度が人間の認識能力を超える程度にまで発展した．また、人間が自然に顔や物体と認識できるような写実的画像を人工的に生成する技術（例えば敵対的生成ネットワーク，GAN)も大きく発展した．研究当初，近い将来，機械学習は人間（エキスパート）の判断や意思決定の一部を肩代わりするようになると予測していた．

深層学習が社会にとって重要な判断や意思決定の一部を担うようになった場合，深層学習が利用される環境が不安定に変化し、訓練時と同じ環境で利用できない状況においても安定した動作をすることが望ましい。あるいは深層学習そのものを不正利用したり，深層学習の判断や意思決定を不正に捻じ曲げて，不当に利益を得ようとしたりする人間が現れると考えられる。そのような敵対的環境において深層学習を適切に動作させるための技術が必要となる。また深層学習の動作自体が信頼されるためには、その挙動が説明可能である必要がある。このような深層学習の安全性や信頼性を向上させるための研究が望まれていた。

２．研究の目的

深層学習の最大の特徴は、人間に匹敵する画像や音声の認識能力を達成する点にある。一方、機械がこれまで扱ってこなかった認識タスクを機械に肩代わりさせることで、これまで想定してこなかったリスクが発生しうる。本研究では、深層学習特有の安全性の問題として、(1)敵対的事例とその防御,(2)生成モデルによる捏造を防ぐ技術，(3)環境変化に対する安定性，(4)説明性に着目した。これらの問題について、技術的解決を与えることがこの研究の目的である。

3.　研究の方法

(1)敵対的事例に関する研究

① 実世界音声敵対的事例

最先端の音声認識モデルを物理世界で攻撃することができる音声敵対例を生成する手法を提案した。これまでの研究では、生成された敵対的サンプルが認識モデルに直接入力されることを前提としており、実際には物理世界では再生環境からの残響やノイズの影響を受けて、攻撃を行うことができなかった。これに対し、提案手法では、物理世界での再生や録音による音声波形の変化をシミュレートし、その変化を生成プロセスに組み込むことで、ロバストな敵対的サンプルを生成する。評価と試聴実験の結果、我々の手法は人間に気づかれずに音声認識モデルを攻撃できることを示した。この結果は，提案手法によって生成された音声敵対的サンプルが実際の脅威となる可能性を示唆している。この研究は国際会議 IJCAI2019 にて発表した。

② 可逆的敵対的事例

可逆的敵対的事例（RAE）という新しいタイプの敵対的事例を提案した。RAE はユーザが指定した AI モデルでは画像を正しく認識・利用できるが、それ以外の AI モデルでは敵対的事例として認識され、正しく認識できない。RAE を実現するために、敵対的事例、敵対的摂動を復元するための可逆的データ隠蔽、敵対的摂動を除去するための高機能暗号という 3 つの技術を組み合わせた。実験結果から、提案手法は対応する敵対的攻撃手法と同等の攻撃能力を持つつ、原画像と同等の視覚品質を実現できることが示された。この研究は Pattern Recognition 誌に発表した。

③ 深層学習を使ったコンテンツベース画像検索への敵対的攻撃に対する保障付き防御

深層学習を使ったコンテンツベース画像検索において、敵対的事例に対する保証付き防御手法を開発した。コンテンツベース画像検索における保証付き防御のために、クエリや候補画像の周りに AX が存在しないことを保証する新しい頑健性に定義を考案し、CBIR の認証防御が達成されているかどうかを検証する計算量的に扱いやすい検証アルゴリズムを提案した。実験の結果、提案された目的関数は既存の方法よりも CBIR の認証防御を大幅に改善することが示された。この研究は国際会議 WACV2022 に発表した。

④ 人間に自然に見える/聴こえる敵対的事例
ディープラーニングモデルを混乱させることを目的とした攻撃/攻撃者では、ほとんどの研究が、人間が攻撃に気づかないように改変の大きさを制限することに焦点を当てている。一方、自律走行車に対する攻撃では、例えば、小さな昆虫の画像が停止標識に置かれても、ほとんどのドライバーはおかしいと思わないか、見過ごしてしまうだろう。この研究では、このような、ある物体や信号を模した自然に見える摂動を採用することで、分類モデルに対する攻撃/攻撃者を自然に生成する体系的なアプローチを提案した。まず、画像分類器に対する攻撃において、ターゲットモデルを欺くために自然物の外観を持つ画像パッチを生成する生成的敵対ネットワークを採用し、このアプローチの実行可能性を示した。さらに、提案手法を音声領域にも拡張できることを実験的に示し、例えば、鳥のさえずりのように聞こえる摂動を生成して音声分類器を欺くことが可能であることを示した。この研究は国際会議 AAAI2020 にて発表した。

(2) 深層学習モデルの悪用を防ぐ技術

① 深層学種モデルへの透かしによる所有権証明

深層学習モデルの学習にはコストがかかるため、ユーザがモデルを無断で再配布するなどの、悪意のあるユーザによる権利侵害の懸念がある。この研究では、ニューラルネットワークモデルを知的財産として保護する手法を提案した。この問題に対する解決策の一つとして、モデルに電子透かしを埋め込むことで、モデルの所有者がモデルの所有権を外部から確認できるようにする方法が考えられるが、この研究で提案するクエリ修正と呼ばれる新しい攻撃方法を用いると、現在存在するすべての電子透かし方法が、クエリ修正や他の既存の攻撃方法（モデル修正など）に対して脆弱であることが実証された。これらの脆弱性を克服するために、指数型重み付けと名付けた新しい電子透かし手法を提案し、不正なサービスプロバイダによる悪意のある無効化処理の試みに対しても、ニューラルネットワークモデル自体の予測性能を犠牲にすることなく、高い電子透かし検証性能を達成することを実験的に示した。この研究は国際会議 ACM ASIACCS2019 にて発表した。

② 生成モデルによる生成画像の生成元追跡

GAN で生成された画像を、その画像を生成した GAN モデルに帰属させることができる手法を提案した。既存の帰属方法（モデル透かしなど）では、高い帰属性能を得るために、モデル公開前にモデルに対して前処理を行う必要がある。この研究では、モデル公開前の前処理を必要としない、ポストホック帰属手法を提案した。この研究で提案する帰属手法は、帰属する画像がモデルによって生成された場合、潜在的な復元がより良い画像復元を達成できることに基づいて設計されている。実験によって、5 枚以上の GAN 生成画像があれば、前処理を必要とする既存手法とほぼ同等の帰属性能を達成できることを示した。さらに、提案手法は、モデル学習の段階で特別な前処理を必要とする電子透かし(watermark) を用いた GAN 帰属処理に対して、競合する帰属性能を達成することが示された。本手法がモデルの学習段階で特別な前処理を必要とせず、あらゆるモデルに事後的に適用できることを考慮すると、提案手法は十分に優れた帰属性能を有していると結論付けることができる。この研究は国際会議 IJCNN2021 にて発表した。

(3) 環境変化に対する安定性

複数の訓練ドメインを利用して、未知のドメインに汎化するモデルを学習することを目的としたドメイン汎化に取り組んだ。具体的には、モデルの領域汎化能力を向上させるために、敵対的学習によって獲得した「分類が難しい」ドメインで学習データを合成し、増強することを目的とした敵対的データ増強の方法を提案した。この研究では、敵対的ドメイン強化と呼ばれる、新しい敵対的データ補強手法を提案した。これは、画像間の翻訳モデルを用いて、学習データと意味的に異なる、同時に分類が困難な新規ドメインの分布を求めるものである。
従来の研究では、（１）意味的に多様でない分類困難なサンプルしか与えることができなかったため、汎化性能が高くなかった、（２）他のデータ増強手法との組み合わせに制限があった、などの問題点があった。我々は、敵対的データ増強と生成モデルベースの手法の両方の利点を取り入れ、生成モデルに基づき最悪ケース分布を活用することによって、多様な分類困難サンプルを精製でき、かつ、他のドメイン汎化アルゴリズムと容易に組み合わせることができる方法を提案した。提案法を実現するためには、意味的に異なる領域を表す分布からサンプルを抽出することができる生成モデルが必要である。また、意味的に異なる複数の領域に対して、最悪の場合の分布を求めるために、データ生成の分散を制御する必要がある。そこで、我々は、複数の学習用データに対して、意味的に異なるスタイルの混合を指定して画像を生成する条件付き画像間変換モデルも合わせて提案した。この生成モデルを用いることで、提案法によって生成される、意味的に異なるサンプルによる敵対的データ補強に基づく訓練が、画像識別機においてより優れたドメイン汎化性能をもたらすことを、複数のデータセットを用いた性能評価実験によって示すことができた。この研究は IEEE Access 誌にて発表した。

(4) 説明可能 AI

この研究では、「データ X は A、B を持ち、C を持たないので、クラス Y に分類される」という形式のブラックボックス分類器を説明することを目的とした (A、B、C は高レベルの概念である)。課題は、分類器を説明するのに有用な概念の集合、すなわち A、B、C を教師なしで発見しなければならないことである。我々はまず、このような概念を表現し発見するのに適した構造的生成モデルを導入した。次に、データ分布を同時に学習し、特定の概念が分類器の出力に大きな因果的影響を与えるように促す学習プロセスを提案した。また、本手法は、ユーザーの事前知識を容易に統合することができ、概念の高い解釈可能性を誘導することができる。最後に、複数のデータセットを用いて、提案手法が本形態での説明に有用な概念を発見できることを実証した。この研究は国際会議 AAAI2022 にて発表した。

４．研究成果

深層学習技術全般に関わる安全性と信頼性の向上に資する技術を、 (1)敵対的事例とその防御,(2)生成モデルによる捏造を防ぐ技術, (3)環境変化に対する安定性, (4)説明性の観点から提案した。提案技術は、ロバスト最適化、bi-level 最適化、電子透かし、ステガノグラフィ、潜在変数復元、教師なし学習など、さまざまな技術を、深層モデルの安全性向上に利用できる形に改良し、複合的に組み合わせることで達成している。この科学研究費の研究期間を通して、生成モデルの隆盛など、深層学習モデル自体も大きく発展した。開発技術はこういった深層学習モデルの発展とその特性に合わせた安全性と信頼性の向上に資するものである。

５．主な発表論文等

〔雑誌論文〕　計8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　3件）

１．著者名 Thien Q Tran, Kazuto Fukuchi, Youhei Akimoto, Jun Sakuma	４．巻 -
２．論文標題 Unsupervised Causal Binary Concepts Discovery with VAE for Black-box Model Explanation	５．発行年 2022年
３．雑誌名 Proceedings of 36th AAAI conference on artificial intelligence	６．最初と最後の頁 1-9
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 Kazuya Kakizaki, Taiki Miyagawa, Inderjeet Singh, Jun Sakuma	４．巻 -
２．論文標題 Toward Practical Adversarial Attacks on Face Verification Systems	５．発行年 2021年
３．雑誌名 2021 International Conference of the Biometrics Special Interest Group (BIOSIG)	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/BIOSIG52210.2021.9548310	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 Hiromu Yakura, Jun Sakuma	４．巻 -
２．論文標題 Generate (non-software) Bugs to Fool Classifiers	５．発行年 2020年
３．雑誌名 The 28th International Joint Conference on Artificial Intelligence	６．最初と最後の頁 5334-5341
掲載論文のDOI（デジタルオブジェクト識別子） 10.48550/arXiv.1911.08644	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 Thien Q. Tran, Jun Sakuma	４．巻 -
２．論文標題 Robust Audio Adversarial Example for a Physical Attack	５．発行年 2019年
３．雑誌名 25th ACM SIGKDD Conference On Knowledge Discovery And Data Mining	６．最初と最後の頁 2857-2866
掲載論文のDOI（デジタルオブジェクト識別子） 10.24963/ijcai.2019/741	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 Ryota Namba, Jun Sakuma	4．巻 -
2．論文標題 Seasonal-adjustment Based Feature Selection Method for Predicting Epidemic with Large-scale Search Engine Logs	5．発行年 2019年
3．雑誌名 The 2019 ACM Asia Conference on Computer and Communications Security	6．最初と最後の頁 228-240
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3292500.3330766	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 Y Zhe, K Fukuchi, Y Akimoto, J Sakuma	4．巻 10
2．論文標題 Domain generalization via adversarially learned novel domains	5．発行年 2022年
3．雑誌名 IEEE Access 10	6．最初と最後の頁 101855-101868
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2022.3209815	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 S Hirofumi, K Fukuchi, Y Akimoto, J Sakuma	4．巻 -
2．論文標題 Did You Use My GAN to Generate Fake? Post-hoc Attribution of GAN Generated Images via Latent Recovery	5．発行年 2022年
3．雑誌名 2022 International Joint Conference on Neural Networks (IJCNN)	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） 10.1109/IJCNN55064.2022.9892704	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 Kazuya Kakizaki, Kazuto Fukuchi, Jun Sakuma	4．巻 -
2．論文標題 Certified Defense for Content Based Image Retrieval	5．発行年 2023年
3．雑誌名 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)	6．最初と最後の頁 4561-4570
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/WACV56688.2023.00454	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0件

〔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

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