学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　氏名　ZHANG Jiahuan

審査担当者　主　査　教　授　長谷山 美紀
副　査　特任教授　荒木 健治
副　査　特任教授　坂本 雄児
副　査　教　授　土橋 宜典
副　査　教　授　小川 貴弘

学位論文題名

A study on deep learning-based adversarial defense by introducing robust features
（ロバスト特徴の導入による深層学習ベースの敵対的防御に関する研究）

本論文は、深層学習モデルへの攻撃に対してロバストな特徴を獲得するための敵対的防御技術の構築に関する研究成果をまとめたものである。

近年、人工知能分野では深層学習モデルが主流となっている。画像分類や画像の意味内容を説明する画像キャプショニングなど、様々なタスクで広く利用されているが、これらの深層学習に基づく構造は、悪意のあるデータ (以降、敵対的サンプル) からの攻撃に対して極めて脆弱であることが示されている。さらに、敵対的サンプルは、人間の目ではクリーンなサンプルと区別がつかないほど高品質であることから、モデルへの攻撃前に人手で敵対的サンプルを取り除くことは困難である。したがって、悪意のある攻撃者が敵対的サンプルを使って深層学習モデルを攻撃することで、誤った結果が出力されてしまうことから、敵対的サンプルの存在は、モデル構築において大きな障壁となるまでに至っている。以上から、深層学習モデルを敵対的サンプルの攻撃から守るための敵対的防御を目的とした技術の構築が急務となっている。

画像分類に対する敵対的防御の分野では、様々な研究が行われており、主に以下の 4 つのカテゴリに分割される。第一に、元々のデータセットに敵対的な摂動を直接加えて学習する敵対的学習法が挙げられるが、この方法は、防御可能な敵対的サンプルが限られてしまうという問題がある。第二に、敵対的サンプルを検出するアプローチが存在するが、これらの多くは入力画像を正しいクラスに分類することを目的としておらず、入力画像が敵対的サンプルであるかどうかを識別するためのモデル設計に主眼が置かれている。第三に、敵対的な摂動の影響を除去または緩和するための入力データの事前処理が挙げられる。最後に、敵対的攻撃に脆弱な原因をモデルの過学習と捉え、正則化を導入、つまりロバストな学習を可能とするアプローチが近年注目を浴びている。

しかしながら、既存の敵対的防御手法には、対処すべき様々な課題が残存している。1 つ目に、正則化に基づく防御手法は、その有効性と低い計算コストのために普及しているが、具体的にモデル中のどの特徴量が正則化に適しているかについては深く議論されていない。2 つ目として、近年画像とキャプションの関係性に注目したマルチモーダル学習に関する研究が盛んに行われているが、既存の画像キャプショニングモデル内の特徴はロバスト性に欠けることが明らかとなっている。最後に、既存の画像分類モデルの敵対的防御に関する研究は、画像から得られる情報にのみ注目しており、近年注目されているマルチモーダル学習を用いたロバストな特徴を獲得する方法論の

確立には至っていない。

本論文の目的は、画像や画像キャプションを用いた深層学習モデルの敵対的防御を実現することである。この達成のために、本論文では以下の 3 つの側面からロバストな特徴量の獲得手法を提案する。まず、画像分類モデルに特徴正則化機能を備えた Multi-stage Feature Fusion Network を提案する。本手法によって、画像の各チャネル中の大域的な特徴を保持可能となり、従来の敵対的防御に関する手法の改善を実現する。次に、攻撃された画像キャプションに悪影響を受けないモデルを構築するために、力学システム理論に基づいた Recurrent Neural Network (RNN) を構築する。これにより、ロバストなキャプション特徴の獲得を可能とする。最後に、画像キャプションを画像分類モデルに導入したマルチモーダル学習によって、ロバストかつ高精度な画像分類を可能とするモデルを構築する。画像とキャプションという異種データ間のギャップを低減するため制約項の導入により、ロバスト性の向上を実現する。以上より、本論文は、上記 3 つの側面から深層学習モデルにおいてロバストな特徴を獲得することで、敵対的防御における既存の課題を解決可能となる。

本論文の各章の構成は以下の通りである。第 1 章では、本研究の研究背景と目的を説明する。第 2 章では、画像分類モデルに対する敵対的攻撃と防御手法に関する関連研究、および、画像キャプショニングモデルを標的とした敵対的攻撃手法に関する関連研究を紹介する。第 3 章では、ロバストな正則化を可能とする Multi-stage Feature Fusion Network について説明する。第 4 章では、ロバストなキャプション特徴を算出する RNN の構築方法について説明する。第 5 章では、豊富な事前知識を含む事前学習モデルによって生成されたロバストキャプション特徴を画像分類モデルに統合する方法について説明する。最後に、第 6 章では、論文の結論と今後の方向性について議論する。

以上を要約すると、本論文では、画像や画像キャプションなど様々なデータを用いた敵対的防御手法を提案し、その有効性を示した。この貢献は、情報科学分野の発展に寄与するものと認められる。したがって、本論文における著者は、北海道大学博士 (情報科学) の学位を授与される資格を有するものと認める。