【学位論文審査の要旨】

発展を続ける深層学習(DNN)は画像認識、自然言語処理、音声認識など、種々の分野に普及している。画像認識の分野では、特に人間の視覚神経を参考にしたアルゴリズムである畳み込みニューラルネットワーク(CNN)が広く実用レベルで成功を収めている。一方、DNN や CNN モデルでは、敵対的サンプルと呼ばれる、人が知覚できないほどのノイズによる攻撃によって誤動作してしまう問題が指摘されており、結果の信頼性向上のための緊急の課題となっている。本論文の課題の一つは、敵対的サンプルに耐性のある CNN モデルの構築法の提案である。他の課題は、CNN モデルの保護技術の提案である。モデルの学習には一般に膨大なデータ、大きな計算コスト、優れたアルゴリズムが必要である。したがって、学習されたモデルは、貴重な価値を持ち、種々の不正アクセスから保護される必要がある。

上述の課題解決のために、先行研究の多くでは、攻撃方法を仮定して、その仮定のもとでモデルを事前学習する。しかし、仮定しない攻撃に対する防御性能の低下に加え、その事前学習の影響がモデル本来の性能を低下させるという課題があった。このような背景から、本論文では、学習可能な画像変換法という新しい着眼によって、これらの課題を解決するための新しい方法を提案し、それらの有効性を多角的な観点から評価した。

本論文で得られた成果を以下に示す。

（1）敵対的サンプル攻撃に対する新しい防御法として、異なる二種類の量子化法を用いた防御法を提案した。使用する画像の画素値を 1 ビットに限定することによって、付加されたノイズの影響を完全に除去できる条件があることを指摘し、その有効性を実験において検証した。

（2）敵対的サンプル攻撃に対する汎用性の高い防御法として、モデルの学習とテスト用データを、秘密鍵を用いて変換する方法を提案した。提案法は、ノイズが付加されていないデータに対しては本来の分類精度を維持し、かつノイズが付加された場合にもその影響が小さいように設計可能である。三種類の変換法を提案して、最先端の攻撃法の下でその有効性を評価した。

（3）モデルの盗難や不正アクセスからモデルを保護する二つの新しい方法を提案した。提案法は、秘密鍵を用いて画像を変換する方法の拡張であり、第一の方法では、モデルパラメータを入手できても、鍵を持たない不正規ユーザーはモデルの性能を活用できない。第二の方法は、鍵の情報をモデル内に電子透かしとして署名でき、モデルの所有権をその署名から主張することを可能にした。

以上のように、本論文は、敵対的サンプル攻撃と不正使用から CNN モデルをいかに保護するかについて考察を行い、新しい防御法を提案し、その有効性の評価を行ったものである。本論文で提案された新しい CNN モデルの防御法とその展開は、急速に発展を続けている深層学習の安全な利用及び信頼性の向上という課題に対して情報科学的アプローチから新しい視点を与えており、今後のこれらの分野の発展へ大いに寄与することが期待され、

重要な意義があると考えられる。よって博士（情報科学）の学位を授与するに十分な価値を有すると認められる。

（最終試験又は試験の結果）

本学の学位規則に従い最終試験を行った。公開の席上（オンライン）で論文発表を行い、学内外から多数の出席者を得て質疑応答を行った。また論文審査委員により本論文及び関連分野に関する試問を行った。これらを総合的に審査した結果、専門科目についても十分な学力があるものと認め合格と判定した。