

(続紙 1)

京都大学	博士（情報学）	氏名	Thomás Rodrigues Crespo
論文題目	Trustworthy machine learning: bridging adversarial and noise robustness of classifiers via smoothed analysis（信頼可能な機械学習：平滑化分析に基づく敵対的頑健性とノイズ頑健性の統合）		
（論文内容の要旨） 本論文は、敵対的攻撃や入力ノイズに対するニューラルネットワークの頑健性の検討を中心に、信頼可能な機械学習の実現に向けた理論及び数値実験の研究成果をまとめたものである。 第1章では、研究背景を概観するために、まず現在の機械学習の進展と多様な応用例を紹介し、次にその実用化に向けた課題として敵対的攻撃を紹介することで、機械学習における頑健性評価と頑健性向上の重要性を述べている。さらに機械学習における信頼性を多様な観点に分類して紹介して、最後に本論文の構成について述べている。 第2章では、計算アルゴリズムの時間複雑性評価に用いられるSmoothed Analysis（平滑化分析）を参考にすることで、敵対的頑健性（敵対的攻撃に対するニューラルネットワークの応答の頑健性）とノイズ頑健性（入力に含まれるノイズに対するニューラルネットワークの応答の頑健性）というニューラルネットワークの2つの頑健性を統一的に扱う枠組みとしてSmoothed Robustness Analysis（SRA、平滑化頑健性分析）を提案している。まずニューラルネットワークの敵対的頑健性とノイズ頑健性についてそれぞれ説明し、両者の関係解明を目指した先行研究、およびアルゴリズムの平滑化分析を紹介することで、これまで1次元的な対立軸上に位置付けられてきた敵対的頑健性とノイズ頑健性が、相互に補完できる関係として定式化できる可能性を説明している。次に、敵対的頑健性とノイズ頑健性が、それぞれ特定の条件を満たす損失関数に対する最悪評価と平均評価に対応することを数理的に明らかにし、その損失関数の局所平均の最大値を考えることで2つの頑健性が統一的に扱えることを示して平滑化頑健性分析という頑健性分析の新たな枠組みを提案している。さらに具体的に利用する損失関数として、最も単純な0-1損失関数には問題点が存在することを指摘し、その問題点の解決のために、マージン損失の利用を提案している。最後に、そのマージン損失を利用した平滑化頑健性分析から導出される信頼性の指標として、Margin Smoothed Certified Radius（MSCR、マージン平滑化認証半径）を提案している。 第3章では、第2章で提案された平滑化頑健性分析とマージン平滑化認証半径の概念実証として、リプシッツ制約を満たすニューラルネットワークを対象とし、その頑健性を向上させる数値実験について述べている。まずリプシッツ制約を満たすニューラルネットワークの役割や実装に関する近年の研究について説明し、次にリプシッツ制約を満たすネットワークに対してマージン平滑化認証半径の推定値の明示的な表式を導出している。続いてその表式を利用することで、認証半径を学習の目的関数として具体的に取り入れる手法を提案している。この提案手法を全結合ニューラルネットワークによる手書き文字データセットの分類問題、および畳み込みニューラルネットワークによるカラー写真の分類問題に適用して、実際に各ニューラルネットワークの敵対的頑健性とノイズ頑健性の両方が向上することを数値的に示している。 第4章では、第3章で用いた学習方法を改善することで、計算コストを大幅に削減する手法を提案している。第3章で提案した手法では、学習過程においてニューラルネットワークにリプシッツ制約を維持させるために、各層の結合行列を互いに直交化させながら学習を行っていた。しかし、この直交性制約の適用は学習に大きな計算オーバーヘッドを与えるため、近年の実用的な機械学習で利用されるような大規模なニ			

ニューラルネットワークへの適用が困難だという問題があった。この章では、まず、マージン平滑化認証半径を目的関数として取り入れる際に現れるハイパーパラメータを、リプシッツ制約の定数を考慮してスケールし直すことをニューラルネットワークの数理的な考察に基づいて提案し、さらにそのスケールされたハイパーパラメータを動的に調整しながら学習進める手法を提案している。この新たな手法では厳密なリプシッツ制約が不要になるため計算オーバーヘッドの大幅な削減が実現される。最後にこの手法を用いた数値実験を行い、この手法でも敵対的頑健性とノイズ頑健性の両方が向上することを確認している。

第5章では、機械学習の信頼性に関する実用上の目的や倫理的側面を考慮しながら、ニューラルネットワークを用いて顔画像からニホンザルの個体の疼痛を検出する問題に取り組んでいる。まず、倫理的および科学的観点から動物の疼痛検出の重要性について述べ、さらにニホンザルを含む霊長類特有の疼痛検出の問題点として、階層的な社会構造のために動物が疼痛を隠す傾向があること、および倫理的観点から極めて少数の教師データしか利用できないことが述べられている。次に、本章の数値実験で用いる顔画像の詳細やその取得方法について述べ、続いて、画像の前処理と評価指標選択の重要性について述べている。最後にニューラルネットワークを用いた数値実験結果を示し、画像からの顔部分抽出において顔輪郭をマスクで抽出する手法の利用が疼痛検出の性能向上に有効であることを報告している。

第6章では、第2章、第3章、第4章、第5章で得られた研究成果の結論が述べられている。また本研究が示唆する今後の発展の可能性について述べられている。

(論文審査の結果の要旨)

本論文は、ニューラルネットワークの頑健性の検討を中心に信頼可能な機械学習の実現に向けた理論及び数値実験の研究成果をまとめたものであり、敵対的攻撃に対するニューラルネットワークの頑健性と入力ノイズに対するニューラルネットワークの頑健性の関係についての理論及び数値実験結果を述べた部分と、ニューラルネットワークを用いて画像からニホンザルの疼痛検出問題に取り組んだ研究をまとめた部分から構成されている。具体的には以下の3つの研究成果を報告している。

第一の研究では、敵対的攻撃に対するニューラルネットワークの頑健性である敵対的頑健性と入力ノイズに対するニューラルネットワークの頑健性であるノイズ頑健性を統一的に扱う「平滑化頑健性分析」という新たな頑健性分析手法を提案した研究成果について報告している。ニューラルネットワークの頑健性に対する適切な評価指標を確立すること、およびその頑健性を向上させることは、ニューラルネットワークの実用化に向けた重要な課題である。これまで敵対的頑健性とノイズ頑健性の2つが主に研究されてきたが、前者は悲観的過ぎる評価を与え、後者は楽観的過ぎる評価を与えることが多かったため、両者を適切に統合する指標の確立が望まれていた。本研究では、ある条件を満たす損失関数の局所平均の最大値を考えることで、この2つの頑健性が統一的に扱えることを示している。敵対的頑健性とノイズ頑健性の統合を目指したこれまでの先行研究では、両者を1次元的な対立軸上に位置付け、その間を補間するアプローチのみが試みられてきていたが、本研究はそれらとは異なり、両者が対立関係ではなく、相互補完する関係として統合できることを示した点で独創的かつ先駆的な研究成果であり評価できる。この分析手法に基づき、実際のニューラルネットワークの学習に適用可能な目的関数の導出にも成功し、その目的関数を用いることで、実際に敵対的頑健性とノイズ頑健性の両方の向上が可能であることを概念検証として示しており、ニューラルネットワークの頑健性に関する基礎付けとして当該分野で重要な貢献をしたと判断できる。

第二の研究は、第一の研究で提案されたニューラルネットワークの頑健性向上手法について、その計算コストを大幅に削減する新たな学習方法を提案した研究である。第一の研究では、ニューラルネットワークの学習時に、各層の結合行列を直交化させながらリプシッツ制約下で学習させる必要があったが、その計算オーバーヘッドが高いため、近年実用的に利用され始めた極めて大規模なニューラルネットワークへの適用は現実的に困難であるという問題があった。これに対し本研究では、利用するニューラルネットワークの目的関数の特性を数理的に考察することで、目的関数に現れるハイパーパラメータを動的に調整すれば、厳密なリプシッツ制約と直交化なしに頑健性を向上させる学習が可能であることを示し、実際にその手法の有効性を数値的に検証することに成功している。数値実験はまだ予備的な段階であり、実用的に利用される極めて大規模なネットワークへの適用はまだ完了していないものの、数理的考察に基づきニューラルネットワークの実用化に向けた重要な進展を達成したものであり、今後の発展の可能性も含めて評価に値する貢献である。

第三の研究は、ニューラルネットワークを用いて画像からニホンザルの疼痛を検出する研究についてまとめたものである。言語で直接痛みを表現できない動物の疼痛の自動検知は、動物の疾患や創傷の早期治療の第一歩として社会的および倫理的観点から重要なものであるが、ニホンザルのような階層的な社会構造を持つ動物種では、階層内地位の維持のために疼痛を隠蔽する傾向が強く、画像から疼痛を検出することは日々動物と接している専門家でも困難であるという問題があった。また倫理的観点から動物が疼痛を持つと思われた場合は早期に治療が行われるため、疼痛状態の画像が極めて少ないという問題もあった。本研究では、これらの問題に対

し、ニホンザルの行動を常時記録している動画像から自動的に顔部分を抽出することで、ニューラルネットワークによって疼痛検出を行う手法について研究したものであり、顔の輪郭部分を厳密に抽出する前処理の利用によって高い検出精度が達成できることを示すことに成功している。ニホンザルを含む霊長類に対してニューラルネットワークを用いて疼痛検知を実現した研究はなく、今後の波及効果も期待され、本研究は当該分野で重要な貢献を果たしたと評価できる。

以上により、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和6年7月22日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。

なお、本論文は、京都大学学位規程第14条第2項に該当するものと判断し、公表に際しては、当面の間、当該論文の全文に代えてその内容を要約したものとすることを認める。