

Evaluating, Understanding, and Enhancing the Adversarial Robustness of Neural Networks: A multi-faceted study on Adversarial Attacks, Latent Space and Neural Architectures

シャシヤンク, コタヤン

<https://hdl.handle.net/2324/7396255>

出版情報 : Kyushu University, 2025, 博士 (情報科学) , 課程博士
バージョン :
権利関係 :



氏 名	SHASHANK KOTYAN (シャシャンク コタヤン)
論 文 名	Evaluating, Understanding, and Enhancing the Adversarial Robustness of Neural Networks : A multi-faceted study on Adversarial Attacks, Latent Space and Neural Architectures (ニューラルネットワークの敵対的ロバスト性の評価・理解・向上:敵対的攻撃・潜在空間・ニューラルアーキテクチャに関する多面的研究)
論文調査委員	主 査 九州大学 准教授 東藤 大樹 副 査 IBM Principal Research Scientist Pin-Yu Chen 副 査 九州大学 准教授 VARGAS DANILO

論 文 審 査 の 結 果 の 要 旨

本研究は、ニューラルネットワークが直面する敵対的脆弱性という根深い課題に対し、その核心を「評価」「理解」「強化」の三位一体のアプローチで解明し、得られた知見に基づいて具体的な堅牢性向上手法を提案したものである。著者は、まず既存の評価手法の限界を指摘し、性質の異なる攻撃を組み合わせた包括的な評価フレームワークを確立した。次に、これまでブラックボックスであったネットワーク内部の脆弱性メカニズムを解明するため、潜在空間の局所構造を定量化する独創的な手法「 k^* -分布」を開発した。さらに、これらの発見を活用し、堅牢なアーキテクチャ探索や人間の視覚に着想を得たデータ拡張法を考案し、その顕著な有効性を実証した。本研究は、敵対的堅牢性の基礎的理解を深めると同時に、その能力を向上させるための実用的な道筋を示したものであり、人工知能分野の発展に大きく寄与するものと評価できる。よって本論文は博士（情報科学）の学位論文に値するものと認める。