Graduate School of Fundamental Science and Engineering
Waseda University

博 士 論 文 審 査 報 告 書
Doctoral Dissertation Review Report

論　文　題　目
Dissertation Title

A Study of Inner Representations in Deep Neural Networks for Comprehending Network Behaviors Using Persistent Homology

ディープニューラルネットワークの挙動解析を目的としたパーシステントホモロジーを用いたネットワーク内部構造の研究

申　請　者
(Applicant Name)
Satoru WATANABE
渡辺　聡

Department of Computer Science and Communications Engineering　Research on Parallel and Distributed Architecture

July, 2022

深層ニューラルネットワーク（Deep Neural Networks, DNN）は，画像解析，音声認識，テキスト分類など，様々な分野で利用され目覚ましい性能を発揮している．DNN では，学習によって得られた知識を内部表現として保持しており，DNN の内部表現がネットワークの振る舞いを決定する．しかし，DNN の内部表現は未解明であり，DNN の学習過程の制御，出力の解釈が困難となっている．このため，DNN の内部表現を解明することは，DNN の利活用に必要不可欠となっている．本論文は，パーシステントホモロジー（Persistent Homology, PH）を用いて DNN の内部表現を解明し，今後の DNN のさらなる発展に資することを目指したものである．

PH は，位相的データ解析（Topological Data Analysis, TDA）における中心的な手法であり，複雑なデータ構造を幾何学的なモデリングで低次元化し，摂動に対して安定な位相的不変量を抽出する手法である．近年では，脳科学，生命科学，情報通信，ビッグデータ解析，材料科学など様々な分野でその活用がはじまっている．

本論文の貢献は，以下の二つに集約される．

最初の貢献は，「PH による DNN 内部表現の解析手法の開発」である．PH は，トポロジー空間における単体複体（多面体）に対して設計されている．これに対して，DNN の挙動はネットワークパラメータによって変化するため，DNN の外形的構造だけでなく，ネットワークパラメータを含めて PH で表現するための手法開発が必要となる．これに対して本論文では，ネットワーク層内と層間で類似の挙動を示すニューロンを定義することで，DNN における単体複体の構成法を提案すると共に，同手法に数学的証明を加え単体複体の構成方法を定式化した．

二つ目の貢献は，「PH を用いた DNN 内部表現の理解度向上」である．PH を用いることで，DNN の内部学習状態を把握し，過学習か学習不足かを判断できることを示した．また，同手法をネットワークの枝刈りに適用し，一般的な枝刈り手法である Global Magnitude Pruning（GMP）に比較して，95%の枝刈り時に 12%高い精度を得ることができることを示した．以上，PH により DNN 内部表現の理解度向上が可能であることを示した．

以下，各章の概要と成果について説明する．

第 1 章は，序論である．

第 2 章は，DNN を対象に単体複体の構成法を提案している．具体的には，DNN の出力から各入力の貢献量を計算する手法である Deep Taylor Decomposition から着想を得，隣接するニューロン間の類似性を定義すると共に，正規化と層間伝搬により、異なる層のニューロン間での類似性の定義へと拡張し，同期する（類似する）ニューロンからなる部分グラフから単体複体を誘導している．さらに，構築手法の正しさを数学的に証明すると共に，全結合

層，畳み込み層，プーリング層に対して PH 計算方法を定式化している．これは，PH を用いた DNN 内部表現に関する研究の基礎となるものである．

第 3 章は，DNN のネットワークパラメータの違いによる PH の変化について分析している．PH の表現方法である PH ダイアグラムを用い，学習済み DNN から得られる単体複体に出現する hole（穴）数とその安定性を確認している．分析の結果，PH ダイアグラムは DNN の学習対象となる問題難易度によって変化することを明かにしている．すなわち，問題難易度による DNN 内部表現の変化が，PH ダイアグラムに表現されることを示している．

第 4 章は，PH による DNN 内部表現の解析が DNN の過学習検出に応用できることを実験的に示している．一般に過学習は，訓練データと検証データの間での精度損失から検出できるが，学習データ無しで配布される事前学習済みの DNN に対しては，同手法は有効ではない．こうした事前学習済みの DNN に対して，過学習の度合い（汎化性能）を PH により検出する手法（Persistent Homology-based Overfitting Measure, PHOM）を提案している．また，ネットワーク構造の違いによる PHOM の変動を緩和するために，PHOM を正規化する手法（Normalized PHOM）を提案している．CIFAR-10，CIFAR-100 をはじめとするデータセットに適用し，DNN の過学習度合いを示すことができることを確認している．これにより，訓練データを用いることなく学習済み DNN に対して過学習度合いを判断することができる．

第 5 章は，DNN を対象とした PH によるネットワーク枝刈り手法（PH-based Network Pruning, PHNP）を提案している．DNN は膨大な計算資源を消費するため，IoT 機器などの計算能力や消費電力が限られる機器で動作させることが難しい．このため，学習済み DNN モデルに対してネットワーク枝刈りを行い，精度の低下を抑えつつ計算量を削減することが求められる．PHNP では，類似の挙動を示すニューロンの集合を PH により抽出し，ニューロン間の関連が弱い集合のエッジを優先的に枝刈りの対象にしている．すなわち，強い関連で構成されるニューロンの集合をネットワークに残す戦略をとっている．CIFAR-10 を用いた評価により，一般的な枝刈り手法である GMP（global magnitude pruning method）と比較し，12%高い精度で DNN から 95%のエッジを刈り取ることができることを示している．

第 6 章は，本論文のまとめである．

以上を要するに，本論文では，PH に基づく DNN 内部表現の解析手法を提案すると共に，応用として DNN の過学習・学習不足を訓練データを用いることなく判断する手法，さらに枝刈りへの応用を示し，その有効性を評価している．これらの成果は，DNN の内部表現の解析に大いに貢献するものであり，博士（工学）（早稲田大学）の学位論文として価値あるものと認める．

2022 年 7 月

審査員

主査	早稲田大学教授	博士（工学）（早稲田大学）	山名早人
副査	早稲田大学教授	博士（工学）（早稲田大学）	菅原俊治
副査	早稲田大学教授	博士（工学）（北海道大学）	内田真人
副査	早稲田大学教授	博士（工学）（早稲田大学）	笠井裕之