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　李 広

審査担当者　主　査　教　授　小川 貴弘
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　長谷山 美紀

学位論文題名

A study on data-efficient learning and its medical applications
（データエフィシェントラーニングとその医療応用に関する研究）

本論文は、深層学習分野における効率的な学習を目的とするデータエフィシェントラーニングの技術構築とその医療応用に関する研究成果をまとめたものである。

深層学習は、コンピュータビジョン、自然言語処理、音声認識などの様々な領域において急速な発展を遂げている。近年、AlexNet、ResNet、CLIP、ChatGPT といった分野を代表する深層学習モデルが開発されており、これらのモデルは、大規模なデータセットから学習することで、高性能なパフォーマンスを発揮する。しかし、大規模なデータセットを前提とした場合、ストレージ、データ転送、データ前処理といった観点から解決すべき課題が存在する。本論文では、上記の問題点を解決するために、効率的な学習方法を実現するためのデータエフィシェントラーニングの研究に焦点を当てる。データエフィシェントラーニングとは、限られた量のデータで学習しつつ高性能を維持することを目的とする研究分野である。伝統的な機械学習アルゴリズムは、その学習に大規模なデータセットを必要とするが、大量のデータを収集しラベル付けすることは多くの時間を要することから、現実的に不可能であることが多い。データエフィシェントラーニングは、これらの制約を克服し、小規模データセットから効果的に学習する方法を開発することを目指す。

データエフィシェントラーニングの一つとして、転移学習が挙げられる。転移学習では、大規模なデータセットで事前に訓練されたモデルを、より小さな目標データセットで微調整する。大規模なデータセットから学んだ知識を活用することにより、モデルはより少ない学習データ数で新しいタスクに迅速に適応することが可能となる。さらに、半教師あり学習や弱教師あり学習等の手法もデータエフィシェントラーニングの一つとして考えられる。半教師あり学習では、ラベル付きデータとラベル無しデータを組み合わせてモデルを学習し、ラベル無しデータはモデルの一般化を改善するための追加情報を提供する。一方、弱教師あり学習は、部分的またはノイズの多い教師情報が含まれるデータに対し、不完全なラベルや弱い注釈から学習することを可能にする。

データエフィシェントラーニングは、限定的なデータからの学習という現実世界の問題を解決する重要性の高い分野である。既存手法は、大規模なデータセットによって生じるいくつかの課題を緩和することができるものの、特定のシナリオに適用する際には本質的な制限が存在する。したがって、これらの制限を解決するために、より強力なデータエフィシェントラーニング手法を探求する必要がある。

本論文の目的は、極端に限られたデータやラベル環境におけるモデルの効率的学習を可能とする技術の構築である。この目標を達成するため、本論文では以下の三つのステージからなる新たなデータエフィシェントラーニング手法を提案する。最初のステージでは、データセットの特性と属性を分析することでデータセットの複雑性を評価する。データセットの複雑性を把握可能とすることで、特定のデータセットに適したモデル構造、訓練戦略、データ拡張技術等を選択することが可能となる。二つ目のステージでは、データセットの蒸留という概念を導入する。データセットの蒸留は、より大きなラベル付きデータセットから知識を活用し、より小さくコンパクトなデータセットに蒸留する。蒸留されたデータセットは、目標タスクにとって不可欠な最も関連性の高い情報を保持する。最後に、三つ目のステージでは、自己教師有り学習をデータエフィシェントラーニングとして高度化する。以上、三つのステージを組み合わせることで、既存の課題を効果的に解決することができる。

本論文の構成は以下の通りである。第 1 章では、本研究の背景と目的を説明する。第 2 章では、データエフィシェントラーニングの関連研究を示し、解決すべき問題を明確にする。第 3 章では、データセットの複雑性評価方法について説明する。第 4 章では、効率的な匿名医療データ共有のためのソフトラベルデータセット蒸留に基づいた圧縮胃画像生成方法について説明する。第 5 章では、胃の X 線画像から識別表現を学習するための自己教師有り学習手法を説明する。第 6 章では、胸部 X 線画像からの COVID-19 検出のための自己教師付き転送学習方法を提示する。第 7 章では、COVID-19 検出精度を向上させるための、自己教師有り学習と自己知識蒸留に基づく新たな方法を提案する。第 8 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、新しいデータエフィシェントラーニング手法を提案し、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。