

学位論文要旨

(Summary of the Doctoral Dissertation)

学位論文題目 (Dissertation Title)	Study on Deep Learning with Pseudo-labeling Mechanism for Chest X-ray Image Diagnosis (胸部X線画像診断のための疑似ラベリング機構付き深層学習に関する研究)
氏名(Name)	Gerdprasert Thanawit

人工知能(AI)は現在、日常活動に組み込まれており、その顕著な応用の一つが医学のためのコンピュータ支援診断(CAD)であり、医療画像に関するセカンドオピニオンを提供することで放射線科医を支援します。その結果、AIは疾病診断の効率を高め、放射線科医の負担を軽減しました。AIの popularity が急増したのは、人間の脳の機能を模倣するアルゴリズムであるディープラーニングの成功によるものです。これにより、AIは複雑なタスクを自律的に学習することが可能になりました。しかし、医療分野では、ラベル付けのタスクに専門家が時間を費やしてアノテーションを行う必要があるため、多くのラベル付きデータを取得するのに苦労しています。また、プライバシーや倫理の問題により、データの収集が困難です。

「半教師あり学習」は、先に述べた問題を解決する候補の一つです。この学習方法は、少数のラベル付きデータと比較的多数のラベルなしデータを使用してAIのパフォーマンスを向上させることに焦点を当てています。よく知られている「半教師あり学習」の技術の一つが疑似ラベリングです。疑似ラベリングの目的は、フィールドの専門家にのみ依存するのではなく、ラベルなしデータにラベリングタスクを行うことです。具体的には、限られた数のラベル付きデータでモデルを訓練してエキスパートモデルまたは教師モデルを作成します。その後、エキスパートモデルがラベルなしデータに「疑似」ラベルを付けます。最終的に、疑似ラベル付きデータが元のラベル付きデータセットに追加され、新たに作成されたデータセットでモデルが再度訓練されます。データの量が増えるとディープラーニングのパフォーマンスが向上するという仮定があるため、疑似ラベル付きデータと元のラベル付きデータの組み合わせで訓練されたモデルは、不十分なラベル付きデータのみで訓練されたモデルよりも優れたパフォーマンスを発揮するはずですが。

疑似ラベリングは、「半教師あり学習」方法としてシンプルかつ効果的な方法です。しかし、疑似ラベリングには多くの課題もあります。たとえば、ラベルなしデータに与えられるアノテーションを受け入れるための最適な閾値を見つけることや、ラベル付きデータが不足しているためにエキスパートモデルがラベルなしデータを適切にアノテートできないことが挙げられます。もう一つの課題は、疑似ラベリングプロセスから発生するバイアスに対処することであり、これはデータセット内の特定のタイプの特徴が存在し、生成されたモデルのトレーニングバイアスを引き起こす場合に生じます。この問題は、不均衡なデータセット、疑似ラベル付きデータの誤ったアノテーション、エキスパートモデルのトレーニング時の特定の顕著な特徴など、さまざまな要因によって引き起こされる可能性があります。

この論文では、医療画像分野における疑似ラベリングメカニズムの設計と強化に焦点を当てました。疑似ラベリングを医療画像診断のCADモデル構築に適用することで、通常はラベル付けコストが高すぎるために廃棄されるラベルなしデータをより有効に活用し、全体的なパフォーマンスを向上させることができると考えています。また、X線画像は一般的に使

用される医療画像であるため、ターゲットデータセットとして胸部X線画像を使用しています。この論文は5章から構成されています。

第1章では、研究の背景を紹介し、基礎を築き、擬似ラベリングと医療画像に関するコアコンセプトを説明します。

第2章では、最初に提案された方法「X線肺炎病変検出におけるディープラーニングと擬似ラベリングを使用した物体検出」を紹介します。この章では、擬似ラベリングのコンセプトを物体検出タスクに適用し、研究の目的は病変領域と病変クラスを同時に特定することです。まず、物体検出に関する基本的な概念、たとえばよく知られているアーキテクチャや評価基準について簡単に紹介します。次に、検出パフォーマンスの安定性を向上させるための反復的な擬似ラベリングメカニズムを導入します。具体的には、擬似ラベル付きデータを徐々に生成し、ラベル付きデータと擬似ラベル付きデータの組み合わせで検出モデルも徐々に訓練します。すべての擬似ラベル付きデータを取得してから一度だけ検出モデルを訓練すると、誤ってラベル付けされた擬似ラベル付きデータがモデルを悪化させる可能性があるため、反復プロセス（段階的な訓練）の実装が採用されています。

第3章では、オブジェクト検出から分類へと話題を変更しました。この章では、「ディープラーニングによる疾病分類と擬似ラベリングによる疾病分類」について、異なる2つのディープラーニングアーキテクチャ、すなわち畳み込みニューラルネットワーク（CNN）とビジョントランスフォーマー（ViT）を使用し、一貫性のある正規化という別の半教師あり学習技術を組み込んだ分類フレームワークの擬似ラベリングを提案します。一貫性のある正規化のコアコンセプトは、擬似ラベル付けされたデータがトレーニングデータセットに含まれるために信頼できるものであるためには、CNNとViTの両方によって得られた予測結果が同じでなければならないということです。さらに、データ拡張を適用することにより、擬似ラベリングメカニズムの堅牢性が実現されます。この章では、それぞれのアーキテクチャと一貫性のある正規化のコンセプトについて詳しく説明するとともに、COVID-19胸部X線画像に実装された実験結果も提供します。

第4章では、アンサンブル学習を使用した強化された擬似ラベリングフレームワークを提案しました。アンサンブルは、あらゆる機械学習技術に適用されてきた一般的な技術ですが、擬似ラベリングにおいて、擬似ラベリングメカニズムの効果を示します。アンサンブル学習は、擬似ラベル付きデータで訓練されたモデルが擬似ラベル付きデータの特定の潜在的特徴に偏ってしまう問題を解決することができます。さらに、各タスクの特性を考慮して、オブジェクト検出タスクと分類タスクのためのアンサンブル方法を設計しました。

この論文は、研究の最後の章で結論を出して終わります。私は研究の成果、意義、制限を結論づけてまとめました。また、医療分野における擬似ラベリングによって達成可能なさらなる研究と潜在性についても提案しています。

学 位 論 文 要 旨

(Summary of the Doctoral Dissertation)

学位論文題目 (Dissertation Title)	Study on Deep Learning with Pseudo-labeling Mechanism for Chest X-ray Image Diagnosis (胸部 X 線画像診断のための疑似ラベリング機構付き深層学習に関する研究)
氏 名 (Name)	Gerdprasert Thanawit

Artificial intelligence (AI) is now embedded in daily activities and one of the notable applications is Computer-Aided Diagnosis (CAD) for medicine, which aids radiologists by providing a second opinion on medical images. Consequently, AI has enhanced the efficiency of disease diagnoses and alleviated the workload of radiologists. The surge in AI popularity came from the success of deep learning, an algorithm mimicking human brain functions, enabling it to learn intricate tasks autonomously. However, the medical field is struggling to acquire a large number of labeled data because the labeling task requires an expert to spend their time annotating to make training data for deep learning, and also, it is difficult to collect data due to privacy and ethical issues.

Semi-supervised learning is one of the candidates that can solve the previously mentioned problem. Semi-supervised learning focuses on using a limited number of labeled data and a relatively large number of unlabeled data to improve the performance of AI. One of the well-known semi-supervised learning techniques is pseudo-labeling. The aim of pseudo-labeling is to perform the labeling task on the unlabeled data instead of relying only on the field experts. In detail, the model is trained with a limited number of labeled data to create an expert model or a teacher model. Then, the expert model gives “pseudo” labels to unlabeled data. Finally, pseudo-labeled data are added to the original labeled dataset, and the model is trained again with the newly created dataset. Since there is an assumption that deep learning performs better as the number of data increases, the model trained with the combination of the pseudo-labeled and the original labeled data should perform better than a model trained with only inadequate labeled data.

Pseudo-labeling is a simplistic yet effective method as a semi-supervised learning method. However, there are also many challenges regarding pseudo-labeling, such as finding the optimal threshold for accepting annotation given for the unlabeled data and the expert model being unable to annotate unlabeled data properly due to the lack of labeled data. Another challenge is to address the bias that occurs from the pseudo-labeling process, which arises when specific types of traits are in the dataset, and they cause the training bias of the generated model. This problem can be caused by various factors, such as an imbalanced dataset, incorrect annotation of pseudo-labeled data, and specific prominent traits when training an expert model.

In this dissertation, I focused on designing and enhancing the pseudo-labeling mechanism in the medical image field. I believe that applying pseudo-labeling to building the CAD model of medical image diagnosis can improve the overall performance by making better use of the unlabeled data, which are usually discarded since the labeling cost is too expensive. In addition, I also use Chest X-ray images as the target datasets because X-ray images are the commonly used medical images. The dissertation consists of five chapters.

In the first chapter, I introduce the research background, lay the groundwork, and explain the core concept regarding pseudo-labeling and medical images.

In the second chapter, the first proposed method, “Object detection using deep learning with pseudo-labeling on X-ray pneumonia disease area detection,” is presented. In this chapter, the concept of pseudo-labeling is applied to the object detection task, where the study's goal is to locate the disease area and the disease class simultaneously. First, I briefly introduce the fundamentals regarding object detection, such as the well-known architectures and evaluation criteria. Then, I introduce an iterative pseudo-labeling mechanism to improve the stability of the detection performance. In detail, the pseudo-labeled data are gradually generated, and the detection model is also gradually trained with the combination of labeled and pseudo-labeled data. If we train the detection model only once after obtaining all the pseudo-labeled data, the incorrectly labeled pseudo-labeled data may deteriorate the model; thus, the implementation of an iterative process (gradual training) is adopted.

In the third chapter, I changed the topic from object detection to classification. In this chapter, “Disease classification using deep learning with pseudo-labeling Disease Classification,” I propose a pseudo-labeling for classification framework that uses two different deep learning architectures, namely convolutional neural network (CNN) and vision transformer (ViT), and also incorporates another semi-supervised learning technique called consistency regularization. The core concept of consistency regularization is that in order for pseudo-labeled data to be reliable to include the training dataset, the prediction results obtained by both CNN and ViT must be the same. In addition, the robustness of the pseudo-labeling mechanism is realized by applying data augmentation. A detailed explanation of each architecture and the concept of consistency regularization is also provided in this chapter, as well as the experimental results implemented on COVID-19 chest X-ray images.

In the fourth chapter, I proposed an enhanced pseudo-labeling framework using ensemble learning. While ensemble has been a common technique applied to any machine learning techniques, in pseudo-labeling, I show the effectiveness of the pseudo-labeling mechanism. Ensemble learning can solve the problem of models trained with pseudo-labeled data tending to be biased toward specific latent features in pseudo-labeled data. In addition, I design ensemble methods for solving object detection tasks and classification tasks, respectively, considering the characteristics of each task.

I finished this dissertation with conclusions in the last chapter of the study. I concluded and summarized the findings, the significance of the studies, and limitations. I also suggest further research and potential that could be achieved by the pseudo-labeling in the medical field.

(様式 9 号)

学位論文審査の結果及び最終試験の結果報告書

山口大学大学院創成科学研究科

氏 名	Gerdprasert Thanawit
審査委員	主 査：間普 真吾
	副 査：多田村 克己
	副 査：中村 秀明
	副 査：田村 慶信
	副 査：藤田 悠介
論文題目	Study on Deep Learning with Pseudo-labeling Mechanism for Chest X-ray Image Diagnosis (胸部 X 線画像診断のための疑似ラベリング機構付き深層学習に関する研究)

【論文審査の結果及び最終試験の結果】

深層学習には通常大量の教師データが必要である。教師データとは、入力データとその正解ラベルのペアをいう。具体例としては、胸部 X 線画像（胸部レントゲン写真）に対する「正常」や「異常」の正解ラベルがある。深層学習は教師データをもとに、入力データに対する正しいラベルを出力する学習が可能であるが、医療などの一部分野では、大量の教師データの収集が困難である。この理由として、プライバシーや倫理的問題があること、また医療はその専門性の高さから、正解ラベルの付与は基本的に医師が行う必要があることが挙げられる。教師データ収集の課題を改善する方法として、半教師あり学習がある。これは限られた数の教師データと大量の正解ラベルなしデータを活用して効果的な学習を行う方式である。本論文では、半教師あり学習の一種である疑似ラベリング手法に着目し、教師データが限られる場合でも、学習に必要な教師データを機械が自ら生成するメカニズムの研究を行っている。教師データ不足は、AI の活用が期待される様々な分野において重要な課題であるため、医療のみならず社会に広く AI が普及するための研究としても価値が高い。

本論文では、年次の健康診断などで大量に撮影され、医師の読影の労力が大きい、胸部 X 線画像を対象とし、以下の 3 つのテーマに関する疑似ラベリング手法を提案し、疑似ラベリングデータを活用した深層学習による診断性能の評価を行っている。① 胸部 X 線画像に写る異常領域を検知する問題において、疑似ラベリングモデルの繰返し学習法を提案している。② 胸部 X 線画像における COVID-19 の陰影の有無を識別する問題において、2 種類の深層学習モデルを組合せ、さらに Consistency 正則化を活用した疑似ラベリング手法を提案している。③ ①と②の方式に対して、それぞれアンサンブル学習による拡張を行っている。

本論文の構成と内容は以下の通りである。

第 1 章では、研究の背景と目的、及び論文の構成について述べている。

(様式 9 号)

第 2 章では、異常領域の検知問題を対象とし、疑似ラベリングの誤りに起因する学習への悪影響が拡散されないよう、信頼性の高い疑似ラベルを次第に増やしていく繰返し学習法を提案した。肺炎検知コンペティション（北米放射線学会）で使用された胸部 X 線画像に対して提案手法を適用したところ、検知性能の向上が示された。

第 3 章では、COVID-19 検知コンペティション（北米放射線学会等）で使用された胸部 X 線画像を対象として、COVID-19 の有無を識別する方式を提案した。データ拡張と Consistency 正則化と呼ばれる方法を組合せた疑似ラベリング手法を適用した結果、多様な画像特徴に対する頑健性が強化され、識別性能の向上が示された。

第 4 章では、第 2 章の検知手法、第 3 章の識別手法の改良を目的とし、それぞれアンサンブル学習を用いた手法を提案した。検知手法では、検知モデルである RetinaNet と YoloV5 が予測する異常領域をそれぞれの信頼度に基づいて合成する手法を提案し性能の改善を示した。識別手法では、識別対象のデータに対して 4 種類のデータ拡張を施し、それぞれを識別することで、4 つの予測結果を得、それらの多数決で最終識別結果を出力する手法を提案した。その結果、識別性能の向上を確認できた。

第 5 章では、本研究の成果をまとめ、結論を述べている。

公聴会には 22 名の参加があり、活発な議論が行われた。主な質疑内容として、

- 領域検知と識別の 2 つの問題を扱っているが両フレームワークの詳細と提案手法のポイントを説明してほしい
- 領域検知の実験結果で疑似ラベルの生成数を提示しているがその意図は何か
- 識別の実験結果で弱いデータ拡張と強いデータ拡張を組合せた学習方式が最も良かった理由は何か
- 畳み込みニューラルネットワークの中で ConvNext が他の方式よりも優れていた理由は何か

等があり、いずれの質問に対しても申請者からの的確な回答がなされた。

以上より、本研究は独創性、信頼性、有効性、実用性とも優れ、博士（工学）の論文に十分値するものと判断した。

論文内容及び審査会、公聴会での質問に対する応答などから、最終試験は合格とした。

なお、主要な関連論文の発表状況は以下の通りである。（関連論文計 4 件）

1. Thanawit Gerdprasert, Shingo Mabu, Shoji Kido, Disease Area Detection for Chest X-ray Image Diagnosis Using Deep Learning with Pseudo Labeling and Ensemble Learning, IEEJ Transactions on Electrical and Electronic Engineering, Volume 18, Issue 11, Pages 1772-1780, 2023 年 09 月発行
2. Thanawit Gerdprasert, Shingo Mabu, Object Detection for Chest X-ray Image Diagnosis Using Deep Learning with Pseudo Labeling, 2021 IEEE 12th International Workshop on Computational Intelligence and Applications, Pages 15-19, 2021 年 11 月発行
3. Thanawit Gerdprasert, Shingo Mabu, Pseudo-Labeling With Contrastive Perturbation Using CNN & ViT for Chest X-ray Classification, 2023 IEEE 13th International Workshop on Computational Intelligence and Applications, Pages 65-69, 2023 年 11 月発行
4. Thanawit Gerdprasert, Shingo Mabu, Combination of Contrastive Perturbation and Ensemble Learning on Pseudo-Labeling Using CNN & ViT for COVID-19 X-ray Classification, 29th International Symposium on Artificial Life and Robotics, Pages 673-677, 2024 年 01 月発行