Form 3

Abstract of Thesis

Name (BOWEN WANG)	
Title	Towards Better Representation and Interpretability for Deep Neural Networks on Visual Tasks (視覚タスクにおけるディープニューラルネットワークのより良い表現と解釈可能性に向けて)

Abstract of Thesis

In recent years, Deep Neural Networks (DNNs) have shown their power over many research fields, and related applications are entering people's daily lives with unstoppable momentum. However, the large number of DNNs' training parameters causes difficulty in learning representation from real-world data efficiently, and the black-box nature harms its explainability. To accommodate actual demands, a DNN model should be adaptive (able to learn meaningful representation for different tasks) and trustworthy (interpretable of its decision). In this thesis, we will show how to design a DNN for better representation, as well as interpret its behavior for reliable artificial intelligence (AI). We first show the effect of designing DNNs according to real needs. A well-designed DNN is adaptive to the target task, owning the ability to learn better representation. We also try to unveil the decision made by DNNs through explainable AI (XAI). By embedding a slot-attention-based XAI module, we find that a DNN model is interpretable, and the learning of representation can be benefited from this interpretability. XAI methods are further extended to find representation in a simple classification task. The found representation is transferred as training data for a complex object detection task, realizing weak supervision. In three different real-world scenarios, we evaluate that our proposal can encourage DNNs to learn better representation and let them be interpretable.

様式 7

論文審査の結果の要旨及び担当者

氏　　名　（ Boweng Wang ）			
論文審査担当者		（職）	氏　　名
	主 査	教授	長原　一
	副 査	教授	八木　康史
	副 査	教授	槇原　靖
	副 査	准教授	早志　英朗
	副 査	准教授	中島　悠太

論文審査の結果の要旨

本論文では、ニューラルネットワークの本質に関連する問題として、データの表現をどのように学習するか、および予測プロセスをどのように解釈するかの2点について論じており、またそれらを組み合わせた新たなニューラルネットワーク学習法について議論している。

第1章では、まず深層ニューラルネットワークの本質である表現学習と、表現学習の柔軟性を確保するためにパラメータ数を増やした結果として生じる深層ニューラルネットワークの解釈の困難性に対する説明可能なAIの2点に関してその現状を述べるとともに、表現学習により説明可能なAIを実現するアプローチと、説明可能なAIの技術によって表現学習を実現するアプローチの二つを考えることで、表現学習と説明可能なAIを組み合わせることの重要性を論じている。

第2章では、本論文の関連研究を紹介している。表現学習に関する研究では、データを0と1で表すハッシュによってデータを表現するアプローチを提案しており、これに関連する研究として深層学習でハッシュによる表現を学習する手法を紹介している。表現学習により説明可能なAIを実現するアプローチでは、ごく少数の学習データから識別器を学習するFew-shot Learningを応用として考えており、これについての関連研究について述べている。また、説明可能なAIの技術によって表現学習を実現するアプローチの関連研究として、医用画像に関連する弱教師あり学習による物体検出の手法を紹介している。最後に、説明可能なAIの関連研究を述べるとともに、本論文の技術的な基盤の一つである手法について詳述している。

第3章では、ハッシュによってデータを表現する手法について述べている。この手法は、博物館の所蔵品を画像により検索するシステムを例としており、博物館で利用されている2種の異なる分類コードを予測するタスクによって深層ニューラルネットワークを学習する。この分類コードは階層的な構造を持っており、2種の分類コードのそれぞれ2階層、計4種のラベルを利用することでよりよいハッシュ表現が得られることを実験的に示している。

第4章では、画像の表現として事前学習によって得られた概念を検出し、その概念を含む領域から抽出された特徴量のみを利用してFew-shot Learningを実現する手法について説明している。事前学習によって得られた概念は、その組み合わせによって任意のクラスを表現することを目指すもので、未知のクラスの画像についても少数の訓練画像からその組み合わせを特定することにより、識別を可能にする。加えて、それぞれの概念に対応する画像中の領域を可視化することにより、識別結果の根拠を提示することができる。

第5章では、胸部レントゲン画像から結節などを検出するアプリケーションを考え、弱教師つき学習により検出器を学習する手法について述べている。この手法ではまず画像中の結節の有無を識別する識別器を学習する。この識別器に対して説明可能なAI手法を適用すると、予測結果に対してどの領域を根拠として識別したかが得られる。結節が存在する画像については、根拠領域は基本的に結節が含まれることから、この根拠領域を学習データとして検出器を学習することにより、結節の検出器が実現できる。

第6章では本論文を総括し、第7章は結論を述べるとともに、今後の展開を示す。

以上、本論文では深層ニューラルネットワークの本質に対して三つの新しいアプローチを論じており、博士（情報科学）の博士論文として価値があるものと認める。