様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　４ 月 １１ 日現在

機関番号：１４４０１
研究種目：若手研究
研究期間：2021～2022
課題番号：２１Ｋ１７７６４

研究課題名（和文）Explainable Artificial Intelligence for Medical Applications

研究課題名（英文）Explainable Artificial Intelligence for Medical Applications

研究代表者
李　良知（LI, LIANGZHI）

大阪大学・データビリティフロンティア機構・特任助教（常勤）

研究者番号：１０８７５５４５

交付決定額（研究期間全体）：（直接経費）　3,600,000円

研究成果の概要（和文）：医療分野において信頼性の高いAIを完全に活用するため、このプロジェクトは説明可能なAIモデルを設計することを目的としています。このモデルは、診断結果に加えて、その判断をサポートする正確で視覚的な説明を提供することができます。このプロジェクトでは、以下のサブトピックを中心に研究を行いました：1.自己注意ベースの分類器。この分類器は、本質的に説明可能な推論を行う能力を持っています。2.説明のサイズを制御するための損失関数。3.同時にpostiveとnegativeの説明を出力するためのコラボレーションサブネットワーク。

結果は、主にICCVやCVPRなどのトップカンファレンスで発表されました。

研究成果の学術的意義や社会的意義
機械の判断に正確な理由を示すことができる技術の確立は、医療分野における説明可能なコンピュータ支援診断（CAD）や、病気の症状を認識する方法や手術を専門家のように実行する方法などのスキルを学生に教えるマシンティーチングシステム、また病気の形態/生理学的状態に関する患者の質問に答えるための医療ビジュアルクエスチョンアンサリング（VQA）など、様々な医療AIアプリケーションの向上につながります。さらに、この研究には一般的なAI研究の改善の可能性があります。

研究成果の概要（英文）：To fully enable trustworthy AI for medicine and healthcare, this project aims to design an explainable AI model that can give diagnosis results along with precise and bifunctional visual explanations to support its decisions. In this project, I mainly studied the following sub-topics towards the goal: 1. A self-attention-based classifier that has the ability to conduct intrinsically-explainable inference; 2. A loss function for controlling the size of explanations. I design a dedicated loss named explanation loss, which is used to control the overall explanation size, region number, etc., of the visual explanations; 3. Collaborating sub-networks to output positive and negative explanations simultaneously.

The results are mainly presented in top conferences like IEEE ICCV and IEEE CVPR.

研究分野：Explainable AI

キーワード：Explainable AI　Computer Vision　Medical Images　Deep Learning　Image Classification　Visual Explanation　Computer-aided Diagnosis　Trustable AI

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

１．研究開始当初の背景

Healthcare and medicine are becoming increasingly interested in how artificial intelligence (AI) can support medical decisions while reducing costs and improving efficiencies [1]. However, due to the black-box nature of AI methods, it is extremely difficult for medical professionals to understand how and why a machine decision has been made. The explainability of AI models is especially important for medical systems, which are risk-sensitive and should give full consideration to human rights. A truly-trustworthy AI model should output not only its decisions but also human-understandable reasons explaining why the decisions are given. To achieve this objective, this project is dedicated to developing an **explainable AI (XAI) method for medicine**.

Currently, there are several XAI approaches available in the computer vision (CV) area, which can provide visual explanations by showing which regions AI models are looking at (in the format of heatmap). However, three problems exist when applying them to medical data (as shown in Fig.1): (a) Most of the existing XAI methods are post-hoc methods, which infer explanations after the decisions have been made by the convolutional neural networks (CNNs) that **are never designed to be explainable in the first place** [2]. (b) The explanations usually cover a large area of the input image [3], which makes it **less precise and less useful**. (c) They are always trying to find the likeness for each possible category [3] and **can only answer the question like "why it is this disease",** rather than the opposite question like "why it is not that disease".

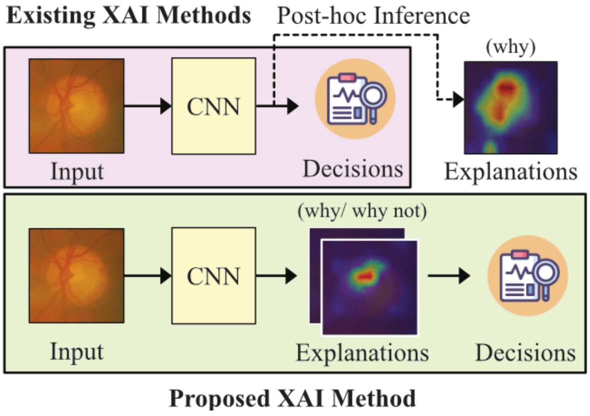


Figure 1. Existing methods (above) and the proposed method (below)

２．研究の目的

As mentioned above, the purpose of this research is to enable intrinsic explainability and controllable explanations for medical AI applications. Research objectives (ROs) include:

- **RO1**: Make an intrinsically-explainable model for common AI tasks, *i.e.*, classification/regression.
- **RO2**: Enable the control over the size of visual explanations.
- **RO3**: Implement the ability to output both positive and negative explanations.

３．研究の方法

This project is divided into three modules, as shown in Fig. 2.

M1: A self-attention-based classifier/regressor that has the ability to conduct intrinsically-explainable inference.

M2: A loss function for controlling the size of explanations.

M3: Collaborating sub-networks to output positive and negative explanations simultaneously.

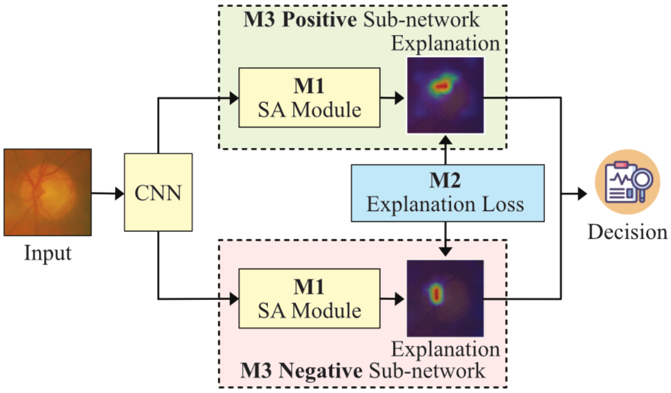


Figure 2. Key modules of the proposed XAI method

References

[1] Andre Esteva, *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, 25(1), pp.24-29, 2019.

[2] Ning Xie, *et al.*, "Explainable deep learning: A field guide for the uninitiated," *arXiv*, 2020.

[3] Pei Wang, *et al.*, "SCOUT: Self-aware Discriminant Counterfactual Explanations," *IEEE CVPR*, 2020.

４．研究成果

(1) Explainable Classifier for Image Recognition

Explainable artificial intelligence has been gaining attention in the past few years. However, most existing methods are based on gradients or intermediate features, which are not directly involved in the decision-making process of the classifier. In this work, I propose a slot attention based classifier called SCOUTER for transparent yet accurate classification. Two major differences from other attention-based methods include: (a) SCOUTER's explanation is involved in the final confidence for each category, offering more intuitive interpretation, and (b) all the categories have their corresponding positive or negative explanation, which tells "why the image is of a certain category" or "why the image is not of a certain category" (as shown in Fig.3). I design a new loss tailored for SCOUTER that controls the model's behavior to switch between positive and negative explanations, as well as the size of explanatory regions. Experimental results

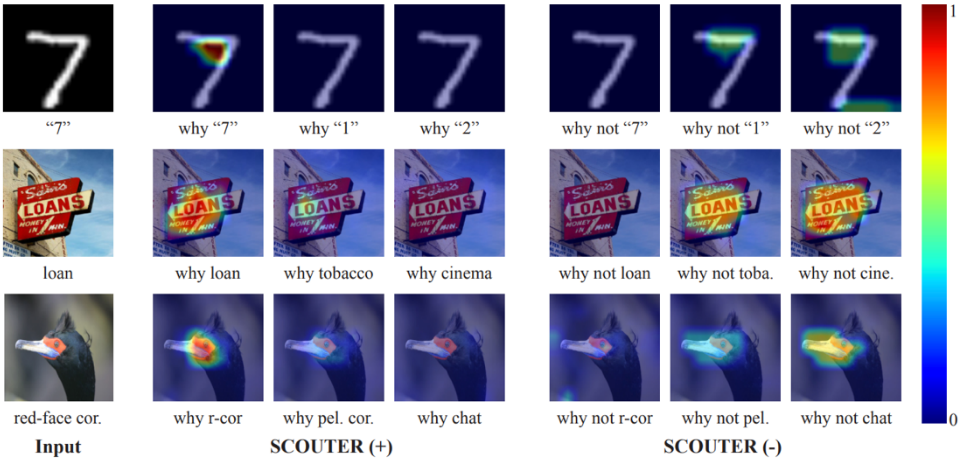


Figure 3. Positive and negative explanations.

show that SCOUTER can give better visual explanations in terms of various metrics while keeping good accuracy on small and medium-sized datasets, as well as some medical tasks (as shown in Fig. 4).

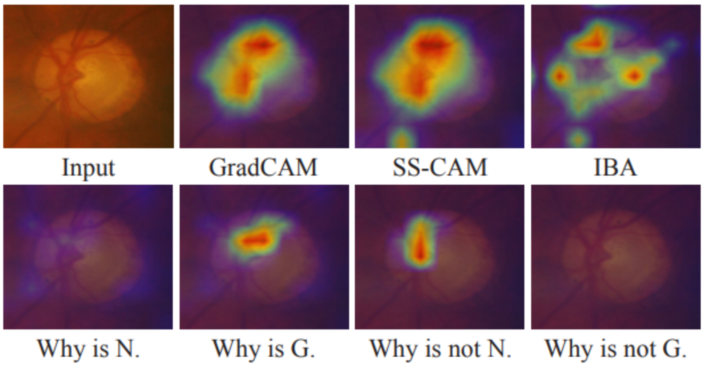


Figure 4. Explanations for a positive sample in a glaucoma diagnosis dataset.

(2) Automated Grading System of Retinal Arterio-Venous Crossing Patterns

The morphological feature of retinal arterio-venous crossing patterns is a valuable source of cardiovascular risk stratification as it directly captures vascular health. Although Scheie's classification, which was proposed in 1953, has been used to grade the severity of arteriolosclerosis as diagnostic criteria, it is not widely used in clinical settings as mastering this grading is challenging as it requires vast experience. In this work, I propose a deep learning approach (as shown in Fig. 5) to replicate a diagnostic process of ophthalmologists while providing a checkpoint to secure explainability to understand the grading process. The proposed pipeline is three-fold to replicate a diagnostic process of ophthalmologists. First, I adopt segmentation and classification models to automatically obtain vessels in a retinal image with the corresponding artery/vein labels and find candidate arterio-venous crossing points. Second, I use a classification model to validate the true crossing point. At last, the grade of severity for the vessel crossings is classified. To better address the problem of label ambiguity and imbalanced label distribution, I propose a new model, named multi-diagnosis team network (MDTNet), in which the sub-models with different structures or different loss functions provide different decisions. MDTNet unifies these diverse theories to give the final decision with high accuracy. This automated grading pipeline was able to validate crossing points with precision and recall of 96.3% and 96.3%, respectively. Among correctly detected crossing points, the kappa value for the agreement between the grading by a retina specialist and the estimated score was 0.85, with an accuracy of 0.92. The numerical results demonstrate that our method can achieve a good performance in both arterio-venous crossing validation and severity grading tasks following the diagnostic process of ophthalmologists. By the proposed models, I could build a pipeline reproducing ophthalmologists' diagnostic process without requiring subjective feature extractions.

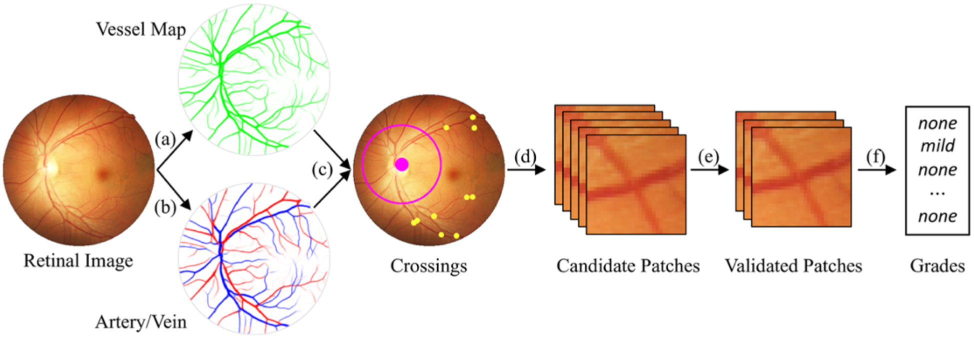
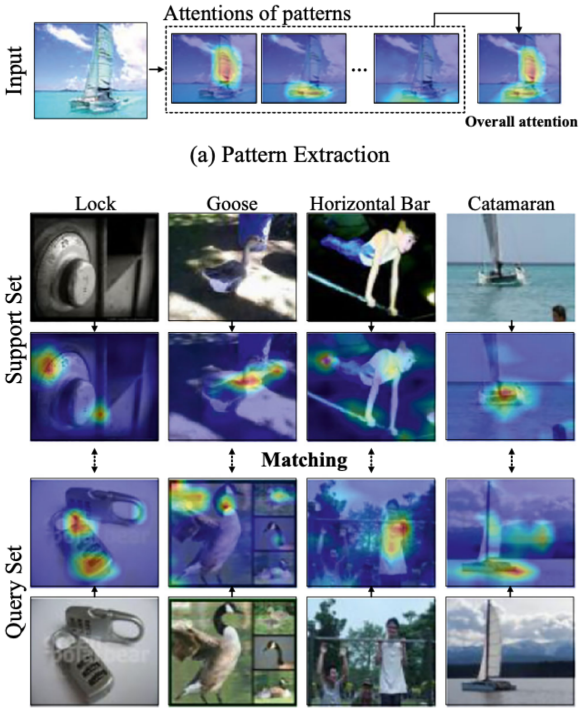


Figure 5. Overall pipeline of the severity grading.

(3) Visually Explainable Few-Shot Image Classification

Few-shot learning (FSL) approaches, mostly neural network-based, assume that pre-trained knowledge can be obtained from base (seen) classes and transferred to novel (unseen) classes. However, the black-box nature of neural networks makes it difficult to understand what is actually transferred, which may hamper FSL application in some risk-sensitive areas. In this work, I reveal a new way (as shown in Fig. 6) to perform FSL for image classification, using a visual representation from the backbone model and patterns generated by a self-attention based explainable module. The representation weighted by patterns only includes a minimum number of distinguishable features and the visualized patterns can serve as an informative hint on the transferred knowledge. On three mainstream datasets, experimental results prove that the proposed method can enable satisfying explainability and achieve high classification results.



(a) Pattern Extraction

(b) Pairwise Matching with overall attention

Figure 6. The proposed visually explainable few-shot learning method.

(4) Explainable Concept Learning for Image Classification

Interpreting and explaining the behavior of deep neural networks is critical for many tasks. Explainable AI provides a way to address this challenge, mostly by providing per-pixel relevance to the decision. Yet, interpreting such explanations may require expert knowledge. Some recent attempts toward interpretability adopt a concept-based framework, giving a higher-level relationship between some concepts and model decisions. In this work, I propose Bottleneck Concept Learner (BotCL), which represents an image solely by the presence/absence of concepts learned through training over the target task without explicit supervision over the concepts. It uses self-supervision and tailored regularizers so that learned concepts can be human-understandable, as shown in Fig. 7. Using some image classification tasks as our testbed, I demonstrate BotCL's potential to rebuild neural networks for better interpretability.

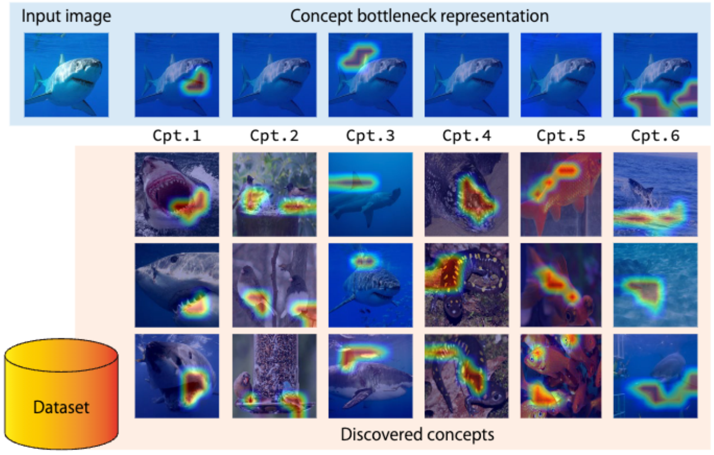


Figure 7. Example concepts extracted by BotCL.

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　2件／うちオープンアクセス　2件）

１．著者名 Li Liangzhi、Verma Manisha、Wang Bowen、Nakashima Yuta、Nagahara Hajime、Kawasaki Ryo	４．巻 2
２．論文標題 Automated grading system of retinal arterio-venous crossing patterns: A deep learning approach replicating ophthalmologist’s diagnostic process of arteriolosclerosis	５．発行年 2023年
３．雑誌名 PLOS Digital Health	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pdig.0000174	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Wang Bowen、Li Liangzhi、Verma Manisha、Nakashima Yuta、Kawasaki Ryo、Nagahara Hajime	４．巻 -
２．論文標題 Match them up: visually explainable few-shot image classification	５．発行年 2022年
３．雑誌名 Applied Intelligence	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10489-022-04072-4	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, Hajime Nagahara
２．発表標題 SCOUTER: Slot Attention-based Classifier for Explainable Image Recognition
３．学会等名 IEEE/CVF International Conference on Computer Vision (ICCV)（国際学会）
４．発表年 2021年

１．発表者名 Bowen Wang, Liangzhi Li, Yuta Nakashima, Hajime Nagahara
２．発表標題 Learning Bottleneck Concepts in Image Classification
３．学会等名 IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR) 2023（国際学会）
４．発表年 2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

＿

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関