東北大学
TOHOKU UNIVERSITY

Press Release

令和 3 年 4 月 27 日

報道機関　各位

東北大学　大学院情報科学研究科
大阪大学　産業科学研究所

説明可能な AI は本当に適切な根拠を示せているのか

AI の説明能力を客観的に評価するための方法論の構築

【発表のポイント】

- AI が下した判断の根拠を明示する手法として、AI が判断を下すに際して参照した過去の類似事例を人間に提示するアプローチが注目されている。
- 類似事例を見つけるための技術的な方法は複数の提案があるが、人間への説明としてどの方法が適切かはわかっていない。
- 予測根拠として最低限満たすべき要件を定式化したところ、非常に人気の高い手法も含め、既存の手法の多くは要件を満たさないことを指摘した。
- 本研究を足がかりに、人間社会に自然に溶け込む AI の研究開発が加速することが期待される。

【概要】

深層学習をはじめとする多くの機械学習手法は、様々な応用分野でその有用性が示されています。一方で機械学習手法は、その判断根拠が不明な点が問題視されています。医療や教育などの分野では特に、「AI がそのような判断をした理由」を知り伝えることが運用上極めて重要です。そのため、最近では「説明可能 AI」という標語の下、AI の判断根拠を示すための研究が盛んに行われています。

説明可能 AI の有力なアプローチのひとつとして、AI が判断のために参照した過去の類似事例を提示する方法があります。東北大学情報科学研究科/理化学研究所の塙一晃研究員、横井祥助教、乾健太郎教授、大阪大学産業科学研究所の原聡准教授らの研究グループは、「類似事例に基づく説明可能 AI」が本当に良い説明を提供しているのかを確かめるため、説明可能 AI が最低限満たすべき要件を定式化し、既存の手法群を再評価しました。研究の結果、既存の手法のいくつかは要件を満たせておらず、要件を満たさない手法群には共通の特徴があることを明ら

かにしました。

研究成果は、現在 AI・機械学習分野でもっとも競争的でインパクトの大きい査読付き国際会議 [1]である The Ninth International Conference on Learning Representations (ICLR 2021) に採択されました。

【詳細な説明】

近年、自動運転や機械翻訳など様々な分野において AI の実用化が進んでいます。一方で特に医療や教育などの分野では、その判断根拠、すなわち「なぜ AI がそのような判断をしたのか」を伝えることが重要です。こうした社会的な要請を受け、近年では機械学習モデルの予測の判断根拠を、人間に説明するための研究が活発に行われています。

判断根拠を説明するための方法は複数考えられますが、一つの有力なアプローチとして「過去のよく似た事例を提示する」というものがあります(図 1)。機械学習モデルを適用する標準的な例として、画像中の鳥の種類を当てるための鳥分類機を構築する状況を考えます。たとえば鳥分類機が、ある鳥画像を“タゲリ”であると分類したとしましょう。有名な鳥ならともかく、タゲリという鳥を知らないユーザーにとっては「これはタゲリです」とだけ示されても納得感に欠けるでしょう。そこで、分類結果とあわせて別のタゲリの画像を提示し「このよく似た画像がタゲリだからこの画像もタゲリだと判断した」というように判断根拠が説明されれば、ユーザーは分類結果により納得できると考えられます。

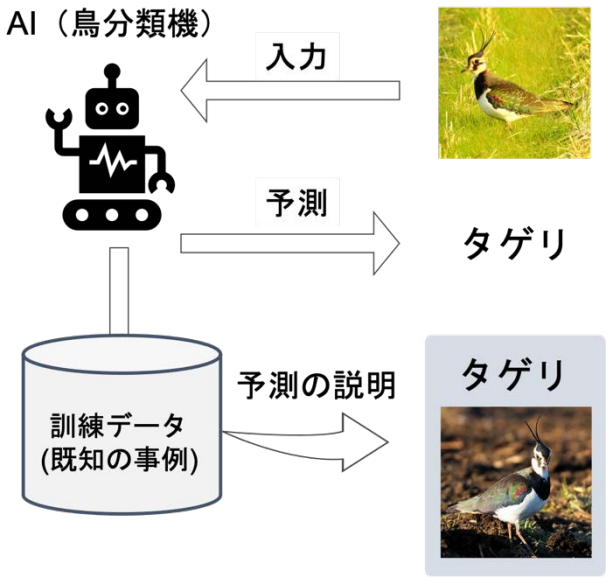

図 1.「過去のよく似た事例を提示する」ことによる判断根拠の説明

[1]https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_artificialintelligence

本研究では「過去のよく似た事例を提示する」説明法が最低限満たすべきいくつかの要件を提案しました。たとえば AI がある画像を猫だと判断した時に、その根拠として「この犬の画像を根拠としてこちらの画像を猫だと判断した」と説明されたとしてもユーザーは納得できないでしょう。我々が提案するひとつめの要件は、「説明に用いる事例は対象としている事例と同じクラスであるべき（ある画像を“猫”だと予測した根拠は“猫”の画像であるべき）」というものです（図 2）。こうした要件に基づき既存の手法を再評価したところ、非常に人気の高い既存の手法も含めた主要な説明法の多くが、これらの最低限とも言える要件を満たさないことがわかりました。また、それらが要件を満たさない原因を数理的に明らかにしました。本研究で得られた知見によって、説明可能 AI の研究が、より社会の要請を満たす方向に推進されることを期待しています。

図 2. 説明法が満たすべき要件の一つ：説明に用いる事例は対象の事例と同じクラスであるべきである

【論文情報】
タイトル: Evaluation of Similarity-based Explanations
著者: Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, Kentaro Inui
掲載誌: Proceedings of The Ninth International Conference on Learning Representations (ICLR 2021)
URL: https://openreview.net/forum?id=9uvhpyQwzM_

【問い合わせ先】
＜研究内容に関すること＞
東北大学　大学院情報科学研究科　教授　乾健太郎
電話　022-795-7091　　E-mail　inui@tohoku.ac.jp

大阪大学　産業科学研究所　准教授　原聡
電話　06-6879-8542　　E-mail　satohara@ar.sanken.osaka-u.ac.jp

＜広報に関すること＞
東北大学　大学院情報科学研究科　広報室 佐藤みどり
電話　022-795-4529　　E-mail　koho@is.tohoku.ac.jp

大阪大学 産業科学研究所 広報室 水野祥子
電話　06-6879-8524　　E-mail　mizuno@sanken.osaka-u.ac.jp