信頼される AI の基盤技術
2021 年度採択研究者

2021 年度
年次報告書

原 聡

大阪大学 産業科学研究所
准教授

機械学習モデルとユーザのコミュニケーション：モデルの説明と修正

§1. 研究成果の概要

研究項目 1「“説明”のための関連性指標の開発と性能検証」については「指標の設計と評価」に取り組んだ。複数の指標を比較検討した結果、「損失関数のパラメータ勾配のコサイン類似度」が最も高い“類推”性能を有することがわかった[1)]。また、研究項目 3 の「指標の計算効率化」についても重点的に研究に取り組んだ。従来は、一晩から数日かけても小規模なモデル・データセットに対してしか関連性指標の計算ができなかった。これに対し、効率化の導入により数倍以上の高速化が可能となり、大規模なモデル・データセットについても関連性指標の計算が可能となった。研究項目 2「“修正”のための関連性指標および効果的なモデル修正方法の開発」については、モデルの再学習による修正に着目し、特に再学習の際のモデルの変化度合いを抑える方法について研究の進展があった。

【代表的な原著論文情報】

1) Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, Kentaro Inui. Evaluation of Similarity-based Explanations. The 9th International Conference on Learning Representations (ICLR'21), 2021.