信頼される AI の基盤技術　　2022 年度
2020 年度採択研究代表者　　年次報告書

西野 正彬

日本電信電話（株）NTT コミュニケーション科学基礎研究所
特別研究員

誤りがないことを保証する検証器つき機械学習の研究

研究成果の概要

機械学習モデルの出力に誤りがないことを保証する、検証器つき機械学習モデルという数理的な枠組みを世界で初めて提案する論文を発表するとともに、検証器を用いることが機械学習モデルの予測性能（汎化誤差）に与える影響を理論的に明らかにした。近年、機械学習技術の大幅な性能向上に伴い、より多様な場面で機械学習技術が活用されるようになりつつある。しかし、機械学習モデルは統計モデルであることから、予測誤りを完全になくすことは難しい。また、誤りが発生できる確率が非常に小さかったとしても、誤りの種類によっては深刻な結果をもたらす可能性がある。検証器つき機械学習は、外部から与えられた「仕様」を機械学習モデルの入出力のペアが満たすかどうかをチェックする検証器を機械学習モデルに取り付けることで、機械学習モデルの入出力が仕様を満たすことを保証可能にする仕組みである。仕様を満たすことを保証することで、AI システムの構築・運用にかかるコストを低減させることができ、より信頼できる AI システムの発展につながると考える。しかし、検証器を機械学習モデルに付与することによってモデルの性能に影響が出ることが予想される。そこで今年度は機械学習モデルに検証器を付加したときに、それが予測誤差に与える影響を理論的に評価した。具体的には、学習時に検証器を用いなかったとしても予測誤差が小さくなることが保証できる条件を提示するとともに、学習時に検証器を用いたならば、予測誤差の理論的上限が悪化しないことを示した。[論文 1]

また、機械学習モデルの検証に関する研究として、自然言語を入力とする深層学習モデルが、入力テキストの微小変化（摂動）に対して予測結果がどのように変化するのかを、数理計画ソルバを用いて厳密に評価する方法を提案した。提案方法を用いることで、既存の評価法よりも高精度に機械学習モデルの頑健性を評価可能であることを示した。[論文 2]

【代表的な原著論文情報】

1) “Generalization Analysis on Learning with a Concurrent Verifier”, In Proc. of the 35th Conference on Advances in Neural Information Processing Systems (NeurIPS), pages 4177-4188, 2022
2) “Robustness Evaluation of Text Classification Models Using Mathematical Optimization and Its Application to Adversarial Training” In Findings of the Association for Computational Linguistics: AACL-IJCNLP, pages 327–333, 2022.