ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名：機械学習モデルの超効率的説明技術の研究	実施時期：2025年4月～2026年3月 所属機関名：千葉工業大学 代表者氏名：吉川 友也

成果概要：

機械学習モデルの予測根拠を少ない計算コストで説明する技術の研究に取り組み、2つの成果を得た。
第一に、入出力のみアクセス可能なブラックボックスモデルを対象として、ネスト構造を持つ入力に対して高・低2レベルの特徴帰属量を一貫性制約つきで同時推定する手法（C2FA）を提案した。本手法はモデルへのクエリ数を従来手法の約1/3に削減しながら同等以上の説明品質を達成し、IJCAI-25に採録された。
第二に、自己説明型モデルの視覚的説明能力をタスク算術（Task Arithmetic）によって予測専用モデルへ追加学習なしで付与する手法を提案し、ABCIの大規模GPU環境を活用して複数データセット（MNIST・CIFAR-10・衛星画像・リモートセンシング等）で検証を行った（arXiv公開済み・学会投稿中）。

成果のポイント：

【成果1】少クエリで一貫した2レベル説明（IJCAI-25採録）

背景と課題: ブラックボックスMLモデルの予測説明には多数のクエリ（入力摂動に対する出力観察）が必要であり、クラウドAPI経由ではコストに直結する。画像分類（MIL）や文書分類など入力が階層的ネスト構造を持つタスクでは、高レベル（画像全体・文）と低レベル（スーパーピクセル・単語）の2段階で説明を提供することが望ましいが、従来手法（LIME, Kernel SHAP等）では2レベルの帰属量を別々に推定するため、高・低レベルで「最も重要」と判定された特徴が矛盾する問題があった。
提案手法（C2FA）: 高レベル特徴帰属量（HiFAs）と低レベル特徴帰属量（LoFAs）の間に成り立つべき一貫性制約を導入し、ADMM（交互方向乗数法）による最適化で両レベルの帰属量を同時推定する。一貫性制約により2レベル間で整合性のとれた説明を生成でき、高・低レベルの摂動クエリを共有できるため必要クエリ数も削減される。
主な実験結果: ABCI 3.0の大規模GPU環境を活用し、画像分類（MIL設定）と言語モデルによるテキスト分類の2タスクで評価。画像分類タスクにおいて、提案手法はクエリ数NL=50で従来手法（LIME等）のNL=150と同等のAUROCを達成し、必要クエリ数を約1/3に削減。高レベル帰属量でも従来手法より少ないクエリ数でNDCGスコア・削除スコアの両指標で最良の結果を示した。定性的にもMILタスクにおける正事例の特定精度や帰属量の視覚的妥当性を確認した。

【成果2】追加学習ゼロの説明能力転移（arXiv公開・学会投稿中）

背景と課題: 成果1がモデルの入出力のみを利用するブラックボックス設定であるのに対し、モデルの重みパラメータにアクセスできる場合にはより根本的なアプローチが可能。近年注目される自己説明型モデル（Self-Explaining Models）は予測と同時に視覚的説明（サリエンシーマップ）を生成できるが、既存の運用中モデルを置き換えるには再学習が必要であり、大量の計算資源とデータが必要になるという課題があった。
提案手法: タスク算術（Task Arithmetic）フレームワークを用いて、自己説明型モデルの重みパラメータから「説明能力に固有のベクトル差分」を抽出し、予測専用モデルのパラメータに加算するだけで説明能力を付与。転移は3ステップで完了：(1) 自己説明型モデルをターゲットデータで学習、(2) 説明能力ベクトルを抽出、(3) 既存モデルに加算。対象モデルのファインチューニングや再学習は一切不要。
主な実験結果: ABCI 3.0の大規模GPU環境を活用し、MNIST・CIFAR-10・衛星画像・リモートセンシングデータセット等で広範な実験を実施。関連性の高いドメイン間では分類精度をほぼ損なわずに視覚的説明品質が向上することを確認。Vision Transformerの拡張アーキテクチャを基盤としており多様なモデルへ適用可能。後処理型手法（SHAP, Grad-CAM等）と比較して、モデルの推論過程により忠実なサリエンシーマップを生成できることも示された。

成果についてより詳細な情報を提供しているWebページ、発表論文などの情報：

1.【採録済み】 Y. Yoshikawa, M. Kimura, R. Shimizu, Y. Saito, "Explaining Black-box Model Predictions via Two-level Nested Feature Attributions with Consistency Property," *Proc. IJCAI-25*, 2025. https://www.ijcai.org/proceedings/2025/0765.pdf
2.【投稿中・arXiv公開】 Y. Yoshikawa, R. Shimizu, T. Kawashima, Y. Saito, "Transferring Visual Explainability of Self-Explaining Models to Prediction-Only Models without Additional Training," *arXiv:2507.04380*, 2025. https://arxiv.org/abs/2507.04380