URL　https://journal.ntt.co.jp/article/36094
DOI　https://doi.org/10.60249/25095005

主 役 登 場

人とAIの連携による価値創造の実現に向けて

山口 真弥 Shinya Yamaguchi

NTTコンピュータ&データサイエンス研究所
革新的コンピューティングアーキテクチャ研究プロジェクト
コンピューティングアルゴリズム研究グループ

近年，大規模言語モデル（LLM）の発展が進み，ChatGPTのようなチャットボットを介して，AI（人工知能）が私たちの生活に急速に浸透し始めています．現在のAIはテキストだけではなく動画像や音声など，多様なモーダルの入出力を受け付けられるようになり，ビジネスや日々の生活を支える幅広いアプリケーションでの活用が進んでいます．今後，次世代のAIはさらなる進化を遂げ，やがて私たちの生活にとってなくてはならない社会インフラとして成長していくことが予想されます．

次世代AIの進化のアプローチの1つとして，現在私たちのグループによって研究開発が進んでいるAIコンステレーション®は，「個性」を持った複数のAIが連携することによって創造性を発揮し，人の高度な知的・創造的業務を支援することをめざしています．これを実現するためには，AIどうしの連携だけでなく，人とAIがどのようにして意思疎通し，合意形成しながら連携するか，という重要な課題を解決しなければなりません．

人とAIが相互に連携するためには，AIが出力の根拠を自分自身で説明できる説明性（Explainability）を提供できることが必要です．しかし，説明性を持つAIの実現はいまだ解決されていない困難な課題です．例えば，現在のAIは出力の根拠を人にとって分かりやすいかたちで説明することを苦手としています．これでは，人はなぜAIがそのように考えたのかを明確に理解できず，追加の指示を与えることが困難になります．また，CoT（Chain-of-Thought：思考の連鎖）のように根拠となる思考を明示的に出力させることができても，AIはしばしば思考過程と整合していない結論を出力してしまう場合があります．したがって，現在のAIが提示する出力の説明性には限界があり，人とAIの正確な連携を妨げています．

私は，これら説明性の課題解決をめざし，研究に取り組んでいます．ここでは，画像を入力とするAIの説明性を実現するXBM（Explanation Bottleneck Model：説明ボトルネックモデル）という研究技術を紹介します．

従来の説明性を持つAI（説明可能AI）は，最終予測の判断根拠として，活性化マップと呼ばれる出力に対して，モデルが入力のどこに反応したのかを大まかに示す可視化や，コンセプトと呼ばれるデータの構成要素（例：色，形状など）を表す，離散的なラベルによる説明形式を採用していました．これらの説明はユーザに一定の理解を与えるものの，表現能力に限界があり，画像内のどのような情報を参照して最終的な予測を行っているのかを正確に理解することは困難でした．

そこで私は，人にとって理解しやすい自然文のテキストによって，直接説明を生成するというアイデアに基づき，新しい説明可能AIであるXBMを提案しました．XBMはマルチモーダル基盤モデルで構成された説明可能AIであり，入力として画像が与えられ，中間出力としてテキストによる説明を生成し，最終出力として画像のクラスラベル（例：猫）を予測する分類モデルです．XBMは画像をマルチモーダルテキストデコーダに入力して説明テキストを生成し，説明テキストから最終クラスラベルを予測する二段階推論によって説明性を担保します．この予測は，通常のクラス分類問題と同様に，クロスエントロピー損失によって学習されます．しかし，XBMを素朴な分類損失だけに頼って訓練すると，テキストデコーダは人にとって意味のないテキストの羅列を生成してしまい，説明性を失ってしまいます．そこで，元の基盤モデルのパラメータを固定した教師モデルから自然な文章を生成し，これによってXBMの出力テキストを罰則する新しい損失を考案しました．つまり，XBMは昔の自分が書いた文章のスタイルを基に，分類タスクの説明に適した語彙を選択するように学習します．この工夫によって，XBMは高い説明性と分類性能を両立しています．また，XBMは説明テキストを編集することで人による介入を受け付け，指示に従って正確に予測結果を修正できる能力を獲得していることも確認しています．これはXBMによる人とAIの連携の可能性を示唆しています．

本技術は画像を入力として検証を行いましたが，音声や言語など，ほかのモーダルの入力データにも適用することが可能です．今後は，XBMで培ったテキストによる説明技術を発展させ，画像，音声，映像，言語など多数の入力形式を扱えるさまざまなAIに説明性を与える研究を進めていきます．そして，これらの説明可能AIを通じて，あらゆるシーンにおいて人とAI，AIとAIをシームレスに連携させること，すなわちAIコンステレーション®の実現をめざします．